The effects of pretesting in econometrics with applications in finance

Danilov, D.L.

Publication date:
2003

Document Version
Publisher's PDF, also known as Version of record

CentER

DMITRY DANILOV

The effects of pretesting in econometrics with applications in finance

TILBURG UNIVERSITY

The effects of pretesting in econometrics with applications in finance

The effects of pretesting in econometrics with applications in finance

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van de rector magnificus, prof.dr. F. A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 7 februari 2003 om 14.15 uur door

Dmitry Leonidovich Danilov

geboren op 13 januari 1969 te Chelyabinsk, Rusland.

Promotor: prof. dr. J.R. Magnus

Acknowledgements

The thesis represents a collection of papers that were written during my participation in the doctoral programme of CentER for Economic Research and I am very much grateful to Department of Econometrics and CentER that provide organizational and financial support in course of this research.

First of all I would like to express my deep gratitude to my supervisor Jan Magnus for his understanding, patience and encouragement during all these years. I would like to thank my colleagues and friends at Tilburg university that create a wonderful spirit of high science, that makes my staying here not only productive but also enjoyable. I am indebted to all teachers of Tilburg University those lectures I attend, who gave me deep understanding of econometrics and economics. I would like to thank all the members of PhD committee, Siem Jan Koopman, Feike Drost, Jim Durbin, Bertrand Melenberg and John Einmahl for their interest in my work.

Last but not least, I am grateful to my wife Olga for her love and support that without doubt was crucial.

Dmitry Danilov
July 2002
Tilburg

Contents

Chapter 1

Introduction

1.1 Pretesting

Pretesting arises in econometrics as soon as we use the same data set both for the model selection and subsequent model estimation. This situation, naturally, can appear not only in economics but in any field where the application of statistical methods is required. It's harmful consequences are however less pronounced in the natural sciences such as physics or astronomy, where a researcher can repeat his measurements as many times as he wants (or, at least as his budget allows). Then it is always possible to have one data set to select a model (hypothesis), and another data set to estimate the selected model. In the non-experimental sciences this is usually not possible, which explains why pretesting is particulary important in the non-experimental sciences. For example, an economist trying to forecast macro-economical indicators for a particular country cannot obtain several instances of this country in order to formulate the model, but will have to work with one set of historical data both for model selection and estimation.

At some occasions authors explicitly pretest in some way, mentioning that this might be a problem. Nevertheless, the common practice is for researchers actually to pretest without mentioning it or perhaps even without realising it. This appears to be the usual situation in applied work. A somewhat more subtle aspect of pretesting is the following. Suppose one author estimates a regression equation using standard statistical techniques, and concludes that a particular regressor has an influence on the dependent variable. This

conclusion is subsequently used by another author, who is lead to either including or excluding this regressor without further testing. This is also an example of pretesting — perhaps we can call it *implicit* pretesting — but it's analysis is more difficult and will not be discussed in the following chapters.

The purpose of our research is to improve the understanding of statistical and economic consequences of pretesting in applied research. In our work we are mostly concerned with implications of pretesting in regression analysis. Typically the applied regression problem takes the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are matrices of regressors, $\boldsymbol{\varepsilon}$ is a random vector of unobservable disturbances, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown parameter vectors. The difference between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is that $\boldsymbol{X}$ always has to be in the model[1], but that the decision whether to use $\boldsymbol{Z}$ or not is taken on the basis of the available data. In the following the columns of $\boldsymbol{X}$ are called *focus* regressors and the columns of $\boldsymbol{Z}$ *auxiliary* regressors.

1.2 The literature

Pretesting has a long history. The earliest work on pretesting is perhaps Berkson (1942). In that paper, the author discusses and criticises routine statistical practice to reject statistical model according to significance level of preliminary test. He proposes to investigate the consequences of preliminary testing in applied research. One of the examples he considers is the linear regression model. Developing Berkson's ideas Bancroft (1944) introduced mathematical framework to analyse the problem of preliminary testing and considers two typical situations where pretesting is relevant: test of homogeneity of variances, and test of a single regression coefficient. He investigates the bias introduced by pretesting. Mosteller (1948) considers the special case $\boldsymbol{x}' = (\boldsymbol{\imath}', \boldsymbol{\imath}')$, $\boldsymbol{z}' = (\boldsymbol{0}', \boldsymbol{\imath}')$, where $\boldsymbol{\imath}$ denotes vector of ones. Thus, Mosteller considers pooling: if $\gamma = 0$ we pool, otherwise we don't pool. In this context, he calculates the mean squared error of the pretest estimator. Huntsberger (1955) also considered the pooling problem. He proposes to use a weighted-average estimator with an arbitrary weight function for

[1]Simplest example of $\boldsymbol{X}$ is an intercept term that normally enters regression equations.

the estimation of the the unknown parameter, arguing, that 'smooth' combinations of the restricted and unrestricted estimators have better properties than standard procedures. The bias and risk for the classical and the smooth pooling pretest estimators were investigated.

The early works of Huntsberger and Mosteller were followed by extensive theoretical investigations of the properties of statistical procedures. One of popular branch of research was connected with investigation of admissibility of statistical procedures. Informally speaking the inadmissible procedure is a procedure whose accuracy can be improved uniformly over the range of relevant parameters. Therefore such a procedure should not be used. However checking for inadmissibility can be a challenging problem and even if estimators are proved to be inadmissible, it is not always possible to find suitable alternatives to them.

The question of admissibility of minimax statistical decision procedures is considered in Wolfowitz (1950) and in Blyth (1951). The later also considered sequential statistical decision rules. Stein (1955) gives necessary and sufficient conditions for admissibility of estimators. Admissibility of estimators with quadratic loss function was investigated in Karlin (1958). Farrel (1968) formulates necessary and sufficient conditions for admissibility of estimators with general strictly convex loss functions. Refinement of Farrel's result lead Brown (1971) to discover a fundamental connection between admissibility of estimators and properties of the diffusion process of special form. Sclove, Morris and Radhakrishan (1972) proved inadmissibility of the pretest estimators for the mean of the multivariate normal distribution and suggested Stein rule estimator as an alternative. For the later problem admissibility issues are discussed in Berger (1976). Another alternative to the usual pretest estimator was proposed in Toro-Vizcarrondo and Wallace (1968), who use mean square error criterion to choose critical regions. Wallace and Ashar (1972) considered the pretest estimator as a weighted-average of the restricted and unrestricted estimators and revealed it's poor performance in the mean squared sense. However no reasonable alternative was proposed. Feldstein (1973) investigated pretesting in regression problems with strongly correlated regressors. In the case of one focus and one auxiliary regressor he investigates performance of the usual pretest and feasible versions of Huntsbegrer's estimators. He concludes that the later behaves better if the value of auxiliary regression parameter is not too large. More traditional approaches to improve properties of the pretest estimator were made in Sawa and Hi-

romatsu (1973) who use minimax regret considerations to set significance level of the pretest estimator. A similar problem was solved in Brook (1976), while Toyoda and Wallace (1976) used minimum average risk criterium. (The review of this early literature is given in Judge and Bock's (1978).)

Sen (1979) compares asymptotical properties of the restricted, unrestricted and pretest maximum likelihood estimators, when pretest is based on likelihood ratio tests. The asymptotic variances of the estimators under various assumptions on parameter space were computed. The author points out that even in general situation the pretest estimator is better than unrestricted estimator in some regions of parameter space, but not uniformly better. Golberger (1981) considers case of the linear regression model where the vector of dependent variables undergos some selection procedure (not all points from original $\boldsymbol{y}$ had been included into regression). He finds that the regression coefficient in this censored regression is a multiple of the uncensored regression coefficient. While previous authors concentrated exclusively on the usual pretest estimator, based on the testing of equality restriction, Thomson and Schmidt (1982) considered the case of the inequality restriction pretest estimator. In particular, exact expression for the risk function was obtained for the case with one focus and one auxiliary regressor. Judge and Bock (1983) give a review of biased estimation. They considered, among others the usual equality and the inequality restriction pretest estimators.

Due to a large interest in the topic, an entire issue of the Journal of Econometrics was devoted to pretesting in 1984. Among others, large contribution was made by Roehrig (1984). He considers a general case of the regression model and investigates mean squared error of the general pretest estimator. He also derives the optimal critical regions imposing priors on γ. However detailed analysis was made only for the case of one auxiliary regressor. Mittelhammer (1984) compares risk of the restricted least squares, the pretest and the Stein-rule estimators when some of regressors were mistakenly omitted from original model. He finds that under this type of model misspecification all alternatives to OLS can be inferior to OLS in terms of prediction risk. This meant that the well established dominance of the Stein-rule estimator over OLS and some other estimators (see Stein (1956) and James and Stein (1961)) is not robust with respect to model misspecification. Therefore it is important to realise that theoretically feasible or even optimal procedures are not always good enough in practice, when the underlying model is actually misspecified. Therefore concepts of admissibility or general risk

optimality should not be the last and decisive arguments in favor of one statistical procedure. Mittelhammer also suggests that estimation of the degree of the misspecification is necessary in choosing a better estimator. Another approach was taken by Zaman (1984), who proposed to avoid using 'popular' model selection procedures due to their inadmissibility and rather use Bayesian procedures. Review of the literature of this period can be find in Giles and Giles (1993).

Magnus and Durbin (1999) derived moments of the general weighted-average least-squares (WALS) estimator. They proved that the problem of finding the optimal WALS estimator in the linear regression model is equivalent to a much simpler *auxiliary* problem. The auxiliary problem was considered in Magnus (2002) under assumption of known variance. Magnus finds an optimal WALS estimator which he calls *neutral Laplace* estimator. While analysis of the pretest estimator was traditionally based on the first two moments, Giles and Srivastava (1993) derive the distribution of the pretest estimator in a regression model with one focus and one auxiliary regressors. Pötscher (1991) derives conditional (on the choice of proper model) asymptotical distributions for the pretest estimator when some kind of general-to-specific procedure is used. This line of research was later continued in Pötscher and Novak (1998), where simulation results were provided. An attempt to apply Pötscher's results to the calculation of confidence and prediction regions was made in Kabaila (1995). Analysis of the asymptotical properties of the regression pretest estimator had also been made also in Zhang (1992). He considered the linear regression model whose order is determined by minimising the generalised final prediction error criterion (see Shibata (1984)). Resulting estimators were proved to be asymptotically unbiased, however their asymptotical variances differ from the variances of the OLS estimators. In Pötscher (2000) the unconditional distributions of the post-model-selection estimators were derived for the special case of backward model selection procedure. A recent review is provided in Magnus (1999).

In view of enormous popularity of regression pretesting we should say that pretesting for estimator of variance of innovations has generated relatively little attention. Toyoda and Walace (1975) consider an F-test for the problem whether to pool two samples in variance estimation. Clark et al. (1987) investigate the risk of the usual preliminary test estimator for the variance of regression disturbances and Wan (1996) examines the case of inequality restrictions.

Most of the papers on pretesting assume normality of innovations. In Giles (1991) the exact risk function of the pretest estimators was derived when disturbances followed general spherically symmetric distribution. The risk performance of the pretest estimator under the balanced loss function was investigated in Ohtani et al. (1997), while Giles (2000) worked with the 'Reflected Normal' loss function.

The problem of pretesting is closely related to the problem of the model selection. However, while the former concentrates on unconditional moments of the estimation (or prediction) given the model selection procedure, as fixed, the later focuses on building the 'correct' model selection procedure. Nevertheless the focus of attention in model selection also lies on some unconditional distributional characteristics such as 'probability of choosing the correct model' etc. The influence of the model selection procedure on the properties of the chosen model are usually omitted. One of the early articles on this topic is Anderson (1962), who considers the problem of choosing the degree in a polynomial regression model. The usual model selection procedures were fairly criticised in many articles. Lovell (1983) investigates performance of several model selection procedures. He notes that usually quoted 5% significance levels are misspecified when several variables are simultaneously tested to find the best model, and proposed an alternative 'rule of thumb' for calculating individual significance levels. The book by Miller (1990) is devoted to model selection in linear regression. He discusses many issues amongst which various methods for selecting the best subset of variables in a linear regression, the effect of selection bias in estimation of regression coefficients as well as regression fundamentals and related topics. Chatfield (1995) frames a wide discussion of different aspects of the model selection. In particular he stressed the fact that predictions and confidence intervals are likely to be narrowed when the estimation phase is preceded by a data driven model building stage. A recent review can be found in George (2000).

Within works on comparative predictive ability we can point out the article of Diebold and Mariano (1995). Under general assumptions they consider test of the null hypothesis of no difference in the accuracy of two competing predictions. West (1996) develops procedures for inference about the moments of smooth functions of out-of-sample predictions and prediction errors and derives asymptotic distributions of relevant predictive statistics. White (2000) considers the case when several null hypothesis about predictive su-

periority are tested simultaneously in order to find the best model (relative to some benchmark model) within several competing ones. For this case he stresses the fact that the proper test statistic is the maximum of several 't-statistics' and exact level of overall test have to be calculated according to the distribution of this maximum. The asymptotic behavior of the test statistic under the null hypothesis was also established and the bootstrap based procedure using this method was applied for several real life examples.

Let us observe, that literature on model selection concentrates on the impact of selection procedures under the null hypothesis. However in practice not only the level but also the power of a test is important. Therefore it is important to know the behavior of the model under different kind of alternatives, this is usually done in pretesting literature. On the other hand it is even more important to realise that model selection actually influences subsequent analysis. We have to know not only the probability of choosing the 'right' model within several competing ones, but also evaluate the accuracy of whole procedure, a usually much more difficult exercise. Finally we should note that a number of classical general-to-specific and specific-to-general model selection procedures are described in textbooks and manuals for well known statistical packages like SPSS or SAS. However none of these sources allow for the fact that preliminary model selection can seriously affect the accuracy of estimation and prediction.

An analysis of the literature would be incomplete without mentioning the data mining methodologies that are closely related to the pretest estimation with several auxiliary regressors. Hoover and Perez (1999) describe a large scale computational experiment for checking various model building strategies. They conclude that a general-to-specific approach works pretty well, and argue in favor of using this approach instead of widely used specific-to-general ones. Hendry (2001) advertises computer-automated general-to-specific procedures and shows that these procedures perform well in Monte Carlo experiments.

1.3 Applications in finance

Investigation of the effects of pretesting in regression analysis has become even more interesting thanks to rapid expansion of the areas where regressions are used. One of the most challenging fields is finance. Originally,

linear regression model appeared in the financial literature as an empirical implication of the capital asset pricing model (CAPM). Black, Jensen and Scholes (1972) are amongst the first who proposed the linear regression model to explain observed assets returns. Fama and MacBeth (1973) introduced cross sectional regression approach. They regressed asset's excess return onto an intercept and 'beta-s' of the CAPM model. Later on the set of explanatory variables was significantly extended and improved. The equity risk premia related variables, such as dividend yield are suggested by Rozeff (1984), while French et al. (1987) proposed default bond premia. Fama and French (1989) suggested to use the interest rates as an explanatory variable, since it affects the overall economic activity and, as a consequence, the stock market activity. Using inflation rate (or other inflation related characteristics) as explanatory factor goes back to Lukas (1986). Industrial production variables are suggested by Balvers et al. (1990) and Chen et al (1986). Price-earnings variables describing how large the stock price is with respect to actual earnings of the company were used in Fama and French (1992). Inspired by development of regression models, Cheng et al. (1990) attempt to forecast the Hong Kong stock price index by multiple regression. However their regression models were not sufficiently powerful to effectively predict the direction of the change in the index. Pesaran and Timmermann (1994) are more successful and demonstrate that the regression model preceded by the variable selection can actually predict movements of the Dow Jones and S&P500 indexes with a sufficient degree of accuracy. These results were enriched and reinforced in Pesaran and Timmermann (1995), where a number of model selection criterions were employed. Problem of forecasting market movements is reconsidered in Granger and Pesaran (2000). They argue that a probability of the fall in the stock market rather than a point stock value is a key element, and propose a way to estimate this probability.

1.4 Contribution of the thesis

Despite the extensive literature on the topic there are still a number of problems to be solved. The fact that pretesting leads to distortion of the accuracy of subsequent estimation is not new. However until now it was not known how large the actual distortions could be. Moreover, the existing literature concentrates almost exclusively on the case of one or two auxiliary regressors, while in practice this number is usually much larger.

In Chapter 2 we fill this gap. We derive the bias, variance and the mean squared error of the pretest estimator under very general assumptions. We generalise Magnus and Durbin's Equivalence Theorem for the case of an arbitrary number of auxiliary regressors and an arbitrary number of preliminary tests. We show that not reporting correct moments can lead to very significant distortions in the accuracy of the estimators even in the case of one auxiliary regressor. We also show that for the case of several auxiliary regressors there are large differences in properties between various model selection procedures. In particular the general-to-specific model selection procedure performs significantly better than specific-to-general. For the specific-to-general procedure not reporting the true moments can lead to unlimited distortion of the accuracy of the estimators. This means that reported variances may have absolutely nothing in common with actual ones. Such an alarming behavior of pretest estimators leads to question the accuracy of pretest estimator as the number of auxiliary regressors grow. We investigate this question for the case when auxiliary regressors are orthogonal in some sense, and find how distortion of accuracy of the pretest estimator grows with the number of regressors. Results of Chapter 2 can be briefly summarised as follows: we find that pretesting in regression analysis can lead to serious problems. Not reporting correct moments of estimator can distort accuracy of the OLS estimators and this effect became stronger as the number of auxiliary regressors grows. In addition the moments of the pretest estimator depends on some unobservable parameters which have to be estimated.

In Chapter 3 we concentrate on two issues. First, we investigate how pretesting affects the accuracy of the one-step-ahead regression forecast. Indeed, if regression forecast was preceded by some model selection, then it's moments would no longer be described by OLS theory. In Chapter 3 we introduce the WALS forecast procedure, that generalises forecasting procedures for regression models selected by preliminary test(s). WALS forecasts (like WALS estimators) not only includes the majority of classical model selection procedures as special cases but also allows for "smooth" combinations of these procedures. We generalise the equivalence theorem of Chapter 2 to describe the unconditional moments of the one-step-ahead WALS forecast. Second, we explore whether pretesting represents a problem in real empirical work. Specifically, we check whether not reporting actual moments will overestimate accuracy of the forecasts. We have chosen to apply our

methods to the case of the stock market forecast considered in Pesaran and Timmermann (1994). For the model selection procedure described by these authors, we calculate unconditional moments and evaluate accuracy of their forecasts. We find that the model selection procedure seriously affects standard errors of the forecasts, and thus reported forecast accuracy is noticeably overestimated. We apply derived theory to the point value forecast and to the probability forecast, introduced in Granger and Pesaran (2000). We also propose several ways to improve the accuracy of the forecast, in particular by orthogonalising the auxiliary regressors. In addition we consider a problem that arises when estimating the moments of the WALS estimator: the natural estimator of these moments is in fact biased and inconsistent.

Chapter 4 contains extensive treatment of the neutral Laplace WALS estimator. Laplace WALS estimator for auxiliary problem was introduced in Magnus (2000) under the assumption of known variance. In Chapter 4 we investigate properties of the WALS estimator in regression problem relaxing this assumption. We propose to estimate the unknown variance by the least-squares estimator of the unrestricted model. We find that the Laplace estimator is admissible, and that it's risk and regret change only marginally when the known variance is replaced by it's estimated value. We also compare the performance of the Laplace and the usual pretest estimator. We find that the Laplace estimator performs better over an practically important range of the parameter. The superiority of the Laplace estimator is more pronounced for small sample sizes.

A short conclusion (Chapter 5) points the way to some unresolved problems.

Chapter 2

On the harm that pretesting does*

2.1 Introduction

In econometrics, due to the non-experimental nature of our discipline, the same data set is commonly used for model selection and for estimation. Standard statistical theory, as developed for the experimental sciences (biology, medicine, physics), is therefore not directly applicable, since the properties of most estimators in econometrics depend not only on the stochastic nature of the selected model, but also on the way this model was selected.

The simplest example of this situation is the standard linear model $y = X\beta + \gamma z + \varepsilon$, where we are uncertain whether to include z or not. The usual procedure is to compute the t-statistic on γ, and then, depending on whether $|t|$ is 'large' or 'small', decide to use the unrestricted or the restricted model. We then estimate β from the selected model. This estimator is a *pretest* estimator, but we commonly report its properties as if estimation had not been preceded by model selection. Thus we report no bias and an incorrect variance.

This is clearly wrong. Our view is *not* that we should avoid pretesting, even though it is well-known that pretest estimators have poor properties, inadmissibility being only one of them. This would be near-impossible in

*Joint paper with J.R. Magnus.

applied work.[1] Our view is simply that we should correctly report the bias and variance (or mean squared error) of the estimators, taking full account of the fact that model selection and estimation are an integrated procedure. This paper attempts to do this.

The literature on pretesting starts with Bancroft's (1944) famous article. Bancroft is mostly concerned with the bias introduced by pretests of homogeneity of variances and pretests of a regression coefficient. He considers the simplest case, in our notation $y = \beta x + \gamma z + \varepsilon$ (one β, one γ), where he wishes to estimate β while being uncertain about whether z should be in the regression or not. He then investigates the bias of the pretest estimator of β. Mosteller (1948) considers the special case $x' = (\imath', \imath')$, $z' = (0', \imath')$, where $\imath$ denotes the vector of ones. Thus, Mosteller considers pooling: if $\gamma = 0$ we pool, otherwise we don't pool. In this context, he calculates the mean squared error of the pretest estimator. Huntsberger (1955) extends Mosteller's paper by explicitly writing the pretest estimator as a (continuous) weighted average of the restricted ($\gamma = 0$) and unrestricted estimator, where the weights are functions of the relevant t-statistic. The fact that the pretest estimator has many undesirable properties is highlighted by Sclove, Morris and Radhakrishnan (1972). Feldstein (1973) is concerned with the problem of estimating β when x and z are highly correlated. He studies the pretest estimator and Huntsberger's weighted average estimator and obtains insights through a simulation experiment. The early literature is discussed in detail in Judge and Bock's (1978) important monograph.

Lovell (1983) asks what will be the true significance level of a t-test after pretesting, and recommends a simple rule-of-thumb. Roehrig (1984) establishes the relationship between the mean squared error of the pretest estimator and the mean squared error of the estimator of the nuisance parameters, a result later generalized by Magnus and Durbin (1999). Mittelhammer (1984) compares the risk functions of several estimators (including the pretest) under model misspecification, and concludes *inter alia* that all alternatives to OLS can be inferior to OLS in terms of prediction risk. The literature of this period is well summarized in Judge and Bock (1983) and in the special issue of the *Journal of Econometrics* (1984), edited by George Judge.

More recently, pretesting has attracted attention in finance, see for example Lo and MacKinlay (1990). Asymptotic aspects are considered in Sen

[1]There are, of course, Bayesian alternatives that avoid model selection. Judge and Bock (1978, 1983) provide a discussion of these. See also Zaman (1984).

(1979), Pötscher (1991), Zhang (1992), and Pötscher and Novak (1998). While most studies, including ours, are confined to the first two moments of the pretest statistics, Giles and Srivastava (1993) derive the distribution of the traditional pretest estimator. Summaries of the latest developments are given in Miller (1990), Giles and Giles (1993), Chatfield (1995), and Magnus (1999).

White (2000), building on work by Diebold and Mariano (1995) and West (1996), provides a method for testing the null hypothesis that the selected model has no predictive superiority over a benchmark model. Different model selection strategies (especially general-to-specific and specific-to-general) are discussed by Hoover and Perez (1999), who favor the general-to-specific procedure. Hendry (2001) advertises computer-automated general-to-specific procedures and claims that these procedures perform well in Monte Carlo experiments. We also find evidence that general-to-specific is preferable over specific-to-general, and find the exact finite sample properties of the two procedures.

In spite of all this literature, we are still far removed from having a fully integrated procedure of model selection and parameter estimation. The current paper attempts to narrow this gap. Our main tool is a generalization of the 'Equivalence Theorem' of Magnus and Durbin (1999). We derive the bias, variance, and mean squared error of the pretest estimator, and show what the error is in not reporting the correct moments. This error can be very substantial. We also show that there can be large differences in underreporting between different model selection procedures. Finally, we ask how the underreporting error increases when the number of auxiliary regressors $z_1, \ldots, z_m$ increases.

The paper contributes to the understanding of the finite sample behavior of the pretest estimator. We only briefly mention asymptotics in our conclusion. The problems do *not* automatically disappear asymptotically, unless one controls the size of the pretests by letting the rejection probabilities tend to zero as the sample size grows, but not too quickly; see Pötscher (1983).

The paper is organized as follows. We define the formal framework and the notation in Section 2. In Section 3 we prove two theorems, which form the basis of the subsequent analysis. Theorem 2 is a generalization of the 'Equivalence Theorem'. In Section 4 we discuss underreporting and its bounds. Section 5 discusses the simplest case, where there is only one auxiliary regressor z. There is only one possible pretest procedure here (using the t-statistic),

and we find, among other things, that in the worst case we report only 13% of the actual pretest mean squared error. In Sections 6 and 7 we address the more difficult case where we have two auxiliary regressors. Then, there is no unique selection procedure. We show, *inter alia*, that there can be large differences between general-to-specific and specific-to-general model selection. Section 8 briefly discusses various extensions and concludes the paper.

2.2 Set-up and notation

The set-up is the same as in Magnus and Durbin (1999) and is briefly summarized. We consider the standard linear regression model

$$y = X\beta + Z\gamma + \varepsilon \tag{2.1}$$

where $y\,(n \times 1)$ is the vector of observations, $X\,(n \times k)$ and $Z\,(n \times m)$ are matrices of nonrandom regressors, $\varepsilon\,(n \times 1)$ is a random vector of unobservable disturbances, and $\beta\,(k \times 1)$ and $\gamma\,(m \times 1)$ are unknown nonrandom parameter vectors. We assume that $k \geq 1$, $m \geq 1$, $n-k-m \geq 1$, that the design matrix $(X : Z)$ has full column-rank $k+m$, and that the disturbances $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$.

The reason for distinguishing between X and Z is that X contains explanatory variables that we want in the model on theoretical or other grounds (irrespective of the found t-values of the β-parameters), while Z contains additional explanatory variables of which we are less certain. Our focus is the estimation of β. Hence the only role for Z is to improve the estimation of β, while γ is a vector of nuisance parameters. The columns of X are called 'focus' regressors, and the columns of Z 'auxiliary' regressors.

We define the matrices

$$M = I_n - X(X'X)^{-1}X' \quad \text{and} \quad Q = (X'X)^{-1}X'Z(Z'MZ)^{-1/2},$$

and the scaled parameter vector $\eta = (Z'MZ)^{1/2}\gamma/\sigma$. The matrix Q can be interpreted as the (scaled) correlation between X and Z. Clearly, $Q = 0$ if and only if Z is orthogonal to X. The least-squares (LS) estimators of β and γ are $b_u = b_r - Q\hat{\theta}$ and $\hat{\gamma} = (Z'MZ)^{-1}Z'My$, where $b_r = (X'X)^{-1}X'y$ and $\hat{\theta} = (Z'MZ)^{1/2}\hat{\gamma}$. The subscripts '$u$' and '$r$' denote 'unrestricted' and 'restricted' (with $\gamma = 0$) respectively. Letting $\hat{\eta} = \hat{\theta}/\sigma$, we see that $\hat{\eta} \sim \mathrm{N}(\eta, I_m)$. Notice that $\hat{\eta}$ is only observable when σ is known, while $\hat{\theta}$ is observable whether σ is known or not.

2.3 The equivalence theorem generalized

Magnus and Durbin (1999) considered the estimation of β in model (1) and proposed a weighted-average least-squares (WALS) estimator of β of the form $b = \lambda b_u + (1-\lambda)b_r$, where $\lambda = \lambda(\hat{\theta}, s_u^2)$ and s_u^2 denotes the estimator for σ^2 in the unrestricted model. This includes the usual pretest estimator as a special case, but only when one restricts the choice of model to the fully restricted and the fully unrestricted case. In this section we prove a generalization of the 'Equivalence Theorem' of Magnus and Durbin, which will allow us to consider not only the unrestricted estimator b_u and the restricted estimator b_r (where *all* γ's are set equal to zero), but also many or all intermediate estimators where *some* of the γ's are set equal to zero. We first state the following preliminary result.

Theorem 1: Let S_i be an $m \times r_i$ selection matrix of rank $r_i \geq 0$, so that $S_i' = (I_{r_i} : 0)$ or a column-permutation thereof. The LS estimators of β and γ under the restriction $S_i'\gamma = 0$ are given by

$$b_{(i)} = b_r - QW_i\hat{\theta}, \qquad c_{(i)} = (Z'MZ)^{-1/2}W_i\hat{\theta},$$

where

$$W_i = I_m - P_i, \qquad P_i = (Z'MZ)^{-1/2}S_i\big(S_i'(Z'MZ)^{-1}S_i\big)^{-1}S_i'(Z'MZ)^{-1/2}$$

are symmetric idempotent $m \times m$ matrices of ranks $m - r_i$ and r_i respectively. (If $r_i = 0$ then $P_i = 0$.) The residual vector is

$$e_{(i)} = y - Xb_{(i)} - Zc_{(i)} = D_i y,$$

where

$$D_i = M - MZ(Z'MZ)^{-1/2}W_i(Z'MZ)^{-1/2}Z'M$$

is a symmetric idempotent matrix of rank $n - k - m + r_i$. The distribution of $b_{(i)}$ is given by

$$b_{(i)} \sim \mathrm{N}\left(\beta + \sigma QP_i\eta,\ \sigma^2\left((X'X)^{-1} + QW_iQ'\right)\right),$$

and the distribution of $s_{(i)}^2 = e_{(i)}'e_{(i)}/(n-k-m+r_i)$ by

$$\frac{(n-k-m+r_i)s_{(i)}^2}{\sigma^2} \sim \chi^2(n-k-m+r_i, \eta'P_i\eta).$$

Proof: Let $X_* = (X : Z)$, $\beta_*' = (\beta', \gamma')$, and $R = (0 : S_i')$. The LS estimator of β_* in the model $y = X_*\beta_* + \varepsilon$ under the restriction $R\beta_* = 0$ is then given by

$$b_* = (X_*'X_*)^{-1}X_*'y - (X_*'X_*)^{-1}R'(R(X_*'X_*)^{-1}R')^{-1}R(X_*'X_*)^{-1}X_*'y.$$

Noting that

$$(X_*'X_*)^{-1} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} = \begin{pmatrix} (X'X)^{-1} + QQ' & -Q(Z'MZ)^{-1/2} \\ -(Z'MZ)^{-1/2}Q' & (Z'MZ)^{-1} \end{pmatrix},$$

and simplifying, the results follow. ||

Several comments are in order. First, we have taken S_i to be a selection matrix such as $S_i' = (0 : I_{r_i})$, so that the restriction $S_i'\gamma = 0$ selects a subset of the γ's to be zero. The theorem, however, only utilizes the fact that S_i has full column-rank. Secondly, if $Q = 0$ (that is, when Z is orthogonal to X) then $b_{(i)} = b_r$ whatever restriction is put on γ, but this is not so for $s^2_{(i)}$. In fact, $s_u^2 \leq s^2_{(i)} \leq s_r^2$, where s_u^2 and s_r^2 denote the estimators for σ^2 in the unrestricted and restricted ($\gamma = 0$) models, respectively. Hence, if $Q = 0$, the pretest estimator is not affected by model selection, but its variance is (see also footnote 3). Thirdly, the normality assumption plays a very minor role in Theorem 1. If we only assume that $\varepsilon \sim (0, \sigma^2 I_n)$, then the expressions for $b_{(i)}$ and $s^2_{(i)}$, the first two moments of $b_{(i)}$, and the first moment of $s^2_{(i)}$ remain the same. Finally, we notice that the partially restricted estimator $b_{(i)}$ is written as a linear function of two vectors b_r and $\hat{\theta}$, which are independent (since $X'y$ and $Z'My$ are independent).[2] Also, $c_{(i)}$ is a linear function of $\hat{\theta}$ only and hence independent of b_r.

If σ^2 is known, then any pretest procedure will use t- and F-statistics which depend on $\hat{\theta}$ only. If σ^2 is not known and estimated by s_u^2, then all t- and F-statistics will depend on $(\hat{\theta}, s_u^2)$. Now, it is a basic result in least-squares theory that s_u^2 is independent of $(b_u, \hat{\gamma})$. It follows that b_r is independent of s_u^2. Hence, b_r will be independent of $(\hat{\theta}, s_u^2)$. Finally, if σ^2 is not known and estimated by $s^2_{(i)}$ corresponding to the selection matrix S_i, then it is no longer true that all t- and F-statistics depend only on $(\hat{\theta}, s_u^2)$. However, they

[2]In fact, even if the observations $y_1, \ldots, y_n$ are not normal and the data-generating process is unknown, b_r and $\hat{\theta}$ will still be uncorrelated, as long as the $\{y_i\}$ are uncorrelated with constant variance (Leeb and Pötscher (2000), Lemma A.1).

still depend only on My, since both $c_{(i)}$ and $e_{(i)}$ are linear functions of My. Hence, the simple fact that b_r and $\hat{\theta}$ are independent implies that all t- and F-statistics used in a pretest procedure, and thus the choice of model, will be independent of b_r.

We are interested in WALS estimators of β, defined as

$$b = \sum_{i=1}^{2^m} \lambda_i b_{(i)}, \tag{2.2}$$

where the sum is taken over all 2^m different models obtained by setting a subset of the γ's equal to zero. Motivated by the previous paragraph, we assume that the weights λ_i satisfy $\lambda_i = \lambda_i(My)$, $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Then,

$$b = b_r - QW\hat{\theta},$$

where

$$W = I_m - P, \qquad P = \sum_{i=1}^{2^m} \lambda_i P_i.$$

Notice that, while P_i and W_i are nonrandom matrices, P and W are random.

Theorem 2 (Equivalence theorem, generalized): Let $b = \sum_i \lambda_i b_{(i)}$, where $\lambda_i = \lambda_i(My)$, $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Then,

$$\mathrm{E}\, b = \beta - \sigma Q\, \mathrm{E}(W\hat{\eta} - \eta), \qquad \mathrm{var}(b) = \sigma^2 \left((X'X)^{-1} + Q\, \mathrm{var}(W\hat{\eta}) Q' \right),$$

and hence

$$\mathrm{MSE}(b) = \sigma^2 \left((X'X)^{-1} + Q\, \mathrm{MSE}(W\hat{\eta}) Q' \right).$$

Proof: Since b_r and My are independent, we have

$$\mathrm{E}(b_r \mid My) = \mathrm{E}(b_r), \qquad \mathrm{var}(b_r \mid My) = \mathrm{var}(b_r).$$

Hence,

$$\begin{aligned} \mathrm{E}(b \mid My) &= \mathrm{E}(b_r \mid My) - Q\, \mathrm{E}(W\hat{\theta} \mid My) \\ &= \mathrm{E}(b_r) - \sigma QW\hat{\eta} = \beta - \sigma Q(W\hat{\eta} - \eta) \end{aligned}$$

and

$$\mathrm{var}(b \mid My) = \mathrm{var}(b_r \mid My) \ = \mathrm{var}(b_r) = \sigma^2 (X'X)^{-1}.$$

The unconditional mean and variance of b and hence its mean squared error follow. ||

This provides a nontrivial generalization, using a simpler proof, of Theorem 2 in Magnus and Durbin (1999). Apparently, the properties of the complicated pretest estimator b of β depend critically on the properties of the less complicated estimator $W\hat{\eta}$ of η.

The restriction that λ_i must depend only on My is a very light one. This allows not only all standard pretest procedures, but also inequality-constrained least squares. Thus, Theorem 2 explains the 'surprising symmetry' found by Thomson and Schmidt (1982, p. 176). The normality assumption plays a stronger role in Theorem 2 than in Theorem 1. Still, if we only assume that $\varepsilon \sim (0, \sigma^2 I_n)$, then Theorem 2 will still hold if the mean and variance of b_r conditional on My are equal to the unconditional mean and variance of b_r.

2.4 Pretesting and underreporting

Theorem 2 shows that if we can find λ_i's such that $W\hat{\eta}$ is an optimal estimator of η, then the same λ_i's will provide an optimal WALS estimator of β. In this paper, however, we are not interested in finding λ_i's such that $W\hat{\eta}$ is an optimal estimator of η. Instead we are interested in the commonly used pretest estimator.

In the idealized context of the linear model $y = X\beta + Z\gamma + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$, we define a *pretest procedure* as a two-step procedure. In step 1 we select the model. In the case $m = 1$ there are two models to choose from: the unrestricted and the restricted (where $\gamma = 0$). In the case $m = 2$ there are four possible models: the unrestricted model, two partially restricted models (one of the two γ's is zero), and the restricted model (both γ's are zero). In general, there are 2^m models to consider in a pretest procedure. We require that the model selection criterion depends on y only through My. In step 2 we estimate the unknown parameters β (and σ^2) from the selected model. This yields the pretest estimators b (and s^2). In a pretest procedure thus defined, the λ_i's are all zero except one which is one.

The mean squared error of the pretest estimator b is, according to Theorem

2,

$$\mathrm{MSE}(b) = \sigma^2 \left((X'X)^{-1} + Q\,\mathrm{MSE}(W\hat{\eta})Q' \right).$$

In applied econometrics practice the same estimator b is selected, but the the effects of pretesting are ignored, the reported bias is zero, and hence the reported MSE equals the reported variance. If we assume that σ^2 is known, then the reported MSE equals

$$\widetilde{\mathrm{MSE}}(b) = \sigma^2 \left((X'X)^{-1} + QWQ' \right),$$

according to Theorem 1, since $W = W_i$ if the i-th model is selected. Notice that $\widetilde{\mathrm{MSE}}(b)$ is random since W is random. Let $\omega'\beta$ be our focus parameter, where ω is an arbitrary nonzero $k \times 1$ vector. In order to compare

$$\mathrm{MSE}(\omega'b) = \sigma^2 \left(\omega'(X'X)^{-1}\omega + \omega' Q\,\mathrm{MSE}(W\hat{\eta})Q'\omega \right) \tag{2.3}$$

with

$$\widetilde{\mathrm{MSE}}(\omega'b) = \sigma^2 \left(\omega'(X'X)^{-1}\omega + \omega' QWQ'\omega \right), \tag{2.4}$$

we define the *underreporting ratio* UR as one minus the ratio of (4) and (3). Thus,

$$\mathrm{UR} = 1 - \frac{\widetilde{\mathrm{MSE}}(\omega'b)}{\mathrm{MSE}(\omega'b)} = \frac{q'(R-W)q}{q'Rq + (1/q_0^2)}, \tag{2.5}$$

where

$$R = R(\eta) = \mathrm{MSE}(W\hat{\eta}), \qquad q = \frac{Q'\omega}{\sqrt{\omega'QQ'\omega}}, \qquad q_0^2 = \frac{\omega'QQ'\omega}{\omega'(X'X)^{-1}\omega}.$$

Notice that $q'q = 1$. The UR is a random variable, since it depends on W, which depends on $\hat{\eta}$. Both the UR and its expectation are unobservable, since they depend on η via $R(\eta)$.

One would expect that the matrix $\mathrm{MSE}(b)$ is at least as large as the matrix $\mathrm{E}(\widetilde{\mathrm{MSE}}(b))$ (in the sense that their difference is positive semidefinite), because pretesting introduces additional noise which is ignored in the reported MSE. Since

$$\mathrm{MSE}(W\hat{\eta}) = \sum_{i=1}^{2^m} \mathrm{E}\,\lambda_i (W_i\hat{\eta} - \eta)(W_i\hat{\eta} - \eta)'$$

and

$$\mathrm{E}(W) = \sum_{i=1}^{2^m} (\mathrm{E}\,\lambda_i) W_i,$$

this is guaranteed if the matrix

$$\sum_{i=1}^{2^m} \mathrm{E}\,\lambda_i\left((W_i\hat{\eta}-\eta)(W_i\hat{\eta}-\eta)'-W_i\right) \tag{2.6}$$

is positive semidefinite. We shall see in the next section that it is possible to devise pretest procedures which do not satisfy this requirement. Such procedures, however, tend to be rather silly. We shall say that a pretest procedure is *viable* if the matrix in (6) is positive semidefinite over the whole parameter space. For any viable pretest procedure, E(UR) is a number between zero and one. When q_0^2 (known to the investigator) tends to zero, then there is no underreporting: E(UR) = 0.[3] But when q_0^2 is large, E(UR) can be close to one.

The $m \times m$ matrix $\mathrm{E}(W)$ is a weighted average of idempotent matrices, and hence is bounded: all its elements are ≤ 1 in absolute value, and all its diagonal elements (and all its eigenvalues) lie in the interval $[0,1]$. In fact,

$$0 \leq \pi_u \leq \xi_j(\mathrm{E}\,W) \leq 1-\pi_r \leq 1 \qquad (j=1,\ldots,m),$$

where $\xi_j(A)$ denotes the j-th eigenvalue of A, π_u is the probability of choosing the unrestricted model ($P_i = 0$), and π_r the probability of choosing the restricted model ($P_i = I_m$).

The E(UR) is a function of q (normalized by $q'q=1$), q_0^2, η, and $Z'MZ$ (and m). Maximizing over q gives the inequality

$$\mathrm{E(UR)} \leq q_0^2 \max_{1\leq j\leq m} \xi_j\left((I_m+q_0^2R)^{-1/2}(R-\mathrm{E}\,W)(I_m+q_0^2R)^{-1/2}\right). \tag{2.7}$$

Then, letting

$$\mathrm{E^*(UR)} = \max_{q,q_0^2} \mathrm{E(UR)},$$

we find, as $q_0^2 \to \infty$,

$$\mathrm{E^*(UR)} = 1 - \min_{1\leq j\leq m} \xi_j(R^{-1/2}(\mathrm{E}\,W)R^{-1/2}) \leq 1 - \frac{\pi_u}{\max_j \xi_j(R)}, \tag{2.8}$$

which depends on η and $Z'MZ$ (and m). We see from (8) that the expected UR can be arbitrarily close to 1 if the mean squared error R fails to be

[3] This happens when $X'Z \to 0$, but also (more generally and less trivially) when $Q'\omega = 0$. In either case $b = b_r$ whatever pretesting we do.

bounded in η. This can not happen when $m = 1$ (unless we *always* choose the restricted model, whatever the value of the observed t-statistic), but it can happen when $m \geq 2$, as we shall see in Section 7.

Finally, since E(UR) depends on $Z'MZ$, we briefly consider the role of this matrix. Without loss of generality, we may scale all z variables so that $z_j'Mz_j = 1$ for all $j = 1, \ldots, m$. In the special case where we can choose the z variables to be 'orthogonal' (in the sense that Mz_i and Mz_j are orthogonal for every $i \neq j$), we have $Z'MZ = I_m$, and major simplifications occur.

Theorem 3: Let $\lambda(x) = 1$ if $|x| > c$ for some $c > 0$, and 0 otherwise. In the special case $Z'MZ = I_m$:

a. W is a diagonal matrix with typical element $w_{jj} = \lambda(\hat{\eta}_j)$;

b. $\mathrm{MSE}(W\hat{\eta}) = V + dd'$, where V is a diagonal $m \times m$ matrix and d an $m \times 1$ vector with typical elements

$$v_{jj} = \mathrm{var}(\lambda(\hat{\eta}_j)\hat{\eta}_j), \qquad d_j = \mathrm{E}(\lambda(\hat{\eta}_j)\hat{\eta}_j - \eta_j);$$

c. The decision whether or not to include z_j in the regression is based exclusively on the t-statistic $\hat{\eta}_j$, and is independent of the selection procedure.

Proof: Using Theorem 1, we have $P_i = S_i(S_i'S_i)^{-1}S_i'$, and, since S_i' is a selection matrix of the form $(I_{r_i} : 0)$ or a column-permutation thereof, it follows that $S_i'S_i = I_{r_i}$ and hence that P_i is a diagonal matrix with r_i ones and $m - r_i$ zeros on the diagonal, and that W_i is a diagonal matrix with $m - r_i$ ones and r_i zeros on the diagonal. Now, also by Theorem 1, $c_{(i)} = W_i\hat{\theta}$ is the estimator of γ under the restriction $S_i'\gamma = 0$. Hence, the estimator of γ_j under this restriction is the j-th component of $c_{(i)}$, which is either 0 (if z_j is excluded from the model) or $\hat{\theta}_j$ (if z_j is included). Thus all models which include z_j as a regressor will have the *same* estimator of γ_j, irrespective which other γ's are estimated. This implies c. Clearly, W is diagonal. The j-th diagonal element w_{jj} is either 0 (if z_j is excluded from the model) or 1 (if z_j is included), that is, $w_{jj} = \lambda(\hat{\eta}_j)$. This implies a. It also implies that the components of $W\hat{\eta}$ are independent of each other, and hence b. follows. ||

Since we shall see that the choice of model selection procedure may matter

a lot for the properties of the estimated focus parameters, it is advisable — if at all possible — to choose the auxiliary regressors such that $Z'MZ = I_m$. This will not only make the the pretest estimator independent of the chosen model selection procedure, but it also allows us to obtain explicit analytical expressions for the moments of the estimator, and it guarantees bounded risk for any value of m. (In the general non-orthogonal case, risk is bounded for $m = 1$, but not necessarily for $m \geq 2$, see Section 7.)

2.5 Underreporting with one nuisance parameter

In the case of one nuisance parameter, the model becomes $y = X\beta + \gamma z + \varepsilon$, where the nuisance parameter γ is a scalar. We have only two models to compare: the unrestricted ($W_1 = 1$, $b_{(1)} = b_u$, $\lambda_1 = \lambda$) and the restricted ($W_2 = 0$, $b_{(2)} = b_r$, $\lambda_2 = 1 - \lambda$). As a result we find

$$b = \lambda b_u + (1 - \lambda) b_r, \qquad W = \lambda,$$

and

$$\mathrm{MSE}(W\hat{\eta}) = \mathrm{MSE}(\lambda\hat{\eta}) = \mathrm{E}(\lambda\hat{\eta} - \eta)^2, \qquad \mathrm{E}\, W = \mathrm{E}\,\lambda.$$

The underreporting ratio is thus

$$\mathrm{UR}(\hat{\eta}, \eta) = \frac{R(\eta) - \lambda(\hat{\eta})}{R(\eta) + (1/q_0^2)},$$

where $\lambda(\hat{\eta}) = 1$ if $|\hat{\eta}| > c$ for some $c > 0$, and 0 otherwise, and

$$R(\eta) = \mathrm{E}(\lambda\hat{\eta} - \eta)^2, \qquad q_0^2 = \frac{(z'X(X'X)^{-1}\omega)^2}{(z'Mz)(\omega'(X'X)^{-1}\omega)}.$$

Assuming again that σ^2 is known and that c is given (say, $c = 1.96$), the λ-function depends only on $\hat{\eta}$, R depends only on η, and hence the UR depends on q_0^2 and $\hat{\eta}$ (both known to the investigator), and η (unknown).

It is easy to see that the larger is $R(\eta)$, the larger is UR. The random variable $\lambda\hat{\eta}$, considered as an estimator of η, thus plays a crucial role in determining the amount of underreporting. We consider its squared bias, variance and MSE in Figure 1.[4]

[4] All results reported in Figures 1–6 are based on exact calculations, with the exception of the locus curve in Figure 3 which relies on numerical techniques.

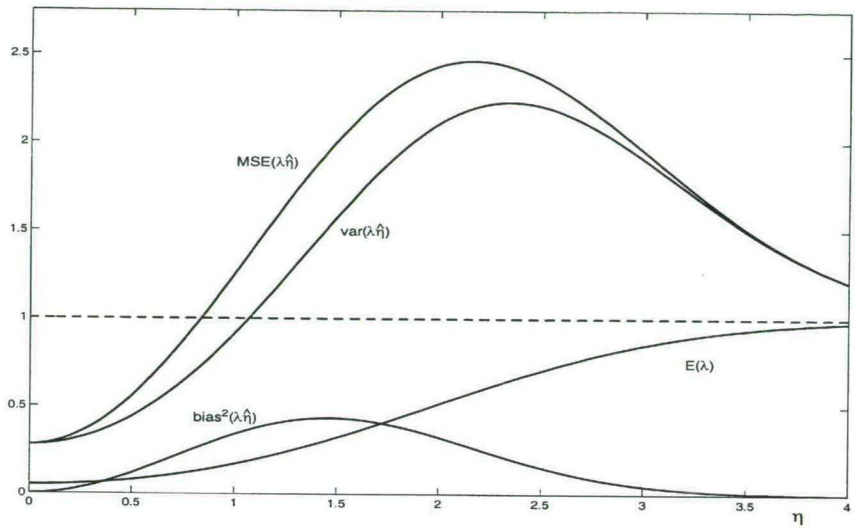

Figure 1. Moments of $\lambda\hat{\eta}$ and λ compared ($m = 1$, $c = 1.96$).

The bias of $\lambda\hat{\eta}$ is negative for $\eta > 0$ and reaches its minimum -0.66 at $\eta = 1.46$. The variance reaches its minimum 0.28 at $\eta = 0$ and its maximum 2.23 at $\eta = 2.34$. The MSE $R(\eta)$ is shaped similarly to the variance. It reaches its minimum at $\eta = 0$ and its maximum 2.46 at $\eta = 2.16$. The variance of $\lambda\hat{\eta}$ is large relative to its bias, suggesting that variance-reduction is more important than bias-reduction.

We also graph the expectation of the reported MSE of $\lambda\hat{\eta}$, that is $\mathrm{E}(\lambda)$, as a function of η for $c = 1.96$, and the MSE of the unrestricted estimator of η, that is $\mathrm{MSE}(\hat{\eta})$ (the dashed line, constant at 1). Since λ only takes the values 0 and 1, $\mathrm{E}(\lambda)$ denotes the probability of choosing the unrestricted model ($\lambda = 1$). But λ also denotes the reported variance (MSE). We see that $\mathrm{E}(\lambda) \equiv \Pr(|\hat{\eta}| > c)$ increases monotonically between 0.05 at $\eta = 0$ and 1 at $\eta = \infty$. Since $\mathrm{MSE}(\lambda\hat{\eta}) \geq \mathrm{E}(\lambda)$, the pretest procedure is viable.[5]

Since λ can only take the values 0 and 1, we can graph the UR for these two values, together with the expected UR and the expectation of λ. This is

[5]However, not all λ-functions lead to a viable procedure. For example, the — admittedly silly — procedure defined by $\lambda = 1$ if $|\hat{\eta}| \leq c$ and 0 otherwise is not viable, since $\mathrm{MSE}(\lambda\hat{\eta}) < \mathrm{E}(\lambda)$ at $\eta = 0$ for any $c > 0$.

done in Figure 2 for the case $q_0^2 = 1$.

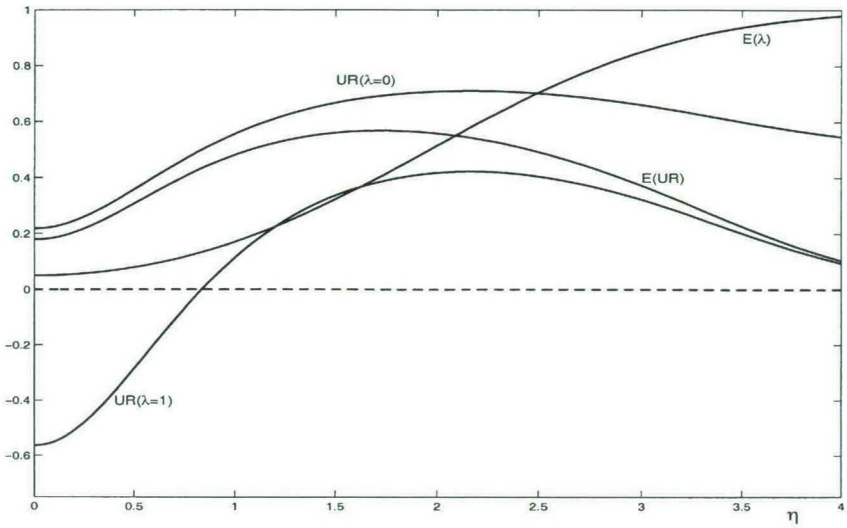

Figure 2. UR (for $\lambda = 0, 1$), E(λ), and E(UR) ($m = 1$, $q_0^2 = 1$, $c = 1.96$).

Figure 2 contains four graphs: the UR at $\lambda = 1$ and at $\lambda = 0$, the expected UR, and E(λ). The graph labeled UR($\lambda = 0$) gives the underreporting ratio when the restricted model is chosen. This function reaches its minimum 0.22 at $\eta = 0$, its maximum 0.71 at $\eta = 2.16$, and approaches $q_0^2/(1 + q_0^2) = 0.5$ as $\eta \to \infty$. Hence, for large values of η, only one half of the actual MSE will be reported when the restricted model is chosen.

Similarly, the graph UR($\lambda = 1$) gives the underreporting ratio when the unrestricted model is chosen. It reaches its minimum -0.56 at $\eta = 0$, its maximum 0.42 at $\eta = 2.16$, and approaches 0 as $\eta \to \infty$. Thus, when η is large and we (correctly) choose the unrestricted model, the UR is zero (no underreporting), but when η is small and we (correctly) choose the restricted model, the UR is still 0.22.

Note that both UR($\lambda = 1$) and UR($\lambda = 0$) reach their maximum at $\eta = 2.16$, where also MSE($\lambda\hat{\eta}$) reaches its maximum. Moreover, the value 2.16 does not depend on q_0^2 (although it does depend on c). Note also that UR($\lambda = 1$) is always smaller than UR($\lambda = 0$), and hence that underreporting is higher if the restricted model is chosen.

When $\lambda = 0$ (and when consequently the restricted model is chosen), the UR always lies between 0 and 1. But when $\lambda = 1$ (unrestricted model), the UR can become negative. This occurs when $|\hat{\eta}|$ is large (> 1.96) but $|\eta|$ is small (< 0.84). In that case the reported MSE is larger than the pretest MSE. The probability that this happens (given by $\mathrm{E}(\lambda)$) is, however, small.

The underreporting ratio $\mathrm{UR}(\lambda = 1)$ does not take account of the probability that the event $\{\lambda = 1\}$ occurs. Neither does $\mathrm{UR}(\lambda = 0)$ take account of the probability that the event $\{\lambda = 0\}$ occurs. In contrast, the expected UR takes account of both probabilities, since it is a weighted average of $\mathrm{UR}(\lambda = 1)$ and $\mathrm{UR}(\lambda = 0)$ with weights $\mathrm{E}(\lambda)$ and $1 - \mathrm{E}(\lambda)$, respectively. We see that $\mathrm{E}(\mathrm{UR})$ is 0.18 at $\eta = 0$, reaches a maximum 0.57 at $\eta = 1.73$, and approaches the curve of $\mathrm{UR}(\lambda = 1)$ as η increases. The $\mathrm{E}(\mathrm{UR})$ varies substantially with η (from 0 to 0.57), indicating that on average the pretest MSE can be 2.3 times the reported MSE ($1/(1 - 0.57) = 2.3$). In contrast to the UR at $\lambda = 0$ or 1, the maximum of $\mathrm{E}(\mathrm{UR})$ *does* depend on q_0. This dependence is analyzed in Figure 3.

In Figure 3 we graph $\mathrm{E}(\mathrm{UR})$ for five different values of q_0^2: 0, 0.1, 1, 10, and ∞. At $q_0^2 = 0$ there is no underreporting and $\mathrm{E}(\mathrm{UR}) = 0$. At $q_0^2 = \infty$, $\mathrm{E}(\mathrm{UR})$ is large; the maximum occurs at $\eta = 0.82$ where $\mathrm{E}(\mathrm{UR}) = 0.87$. This means that the reported variance should be multiplied by about 7.5 in order to obtain the true MSE of the pretest estimator.

Finally, since both UR and $\mathrm{E}(\mathrm{UR})$ depend on η, we also consider the behavior of the underreporting ratio at $\eta = 1$. This is an interesting value, because it is the value of η where the investigator is indifferent between the restricted and the unrestricted model; see Magnus and Durbin (1999, Theorem 1).

Figure 4 shows that the UR at $\eta = 1$ is an increasing function of q_0^2, with $\mathrm{UR} = 0$ at $q_0^2 = 0$. When $q_0^2 \to \infty$, UR approaches 1 when $\lambda = 0$ and 0.20 when $\lambda = 1$, since $R(1) = 1.26$. The expectation of UR approaches 0.86, since $\mathrm{E}(\lambda) = 0.17$ at $\eta = 1$.

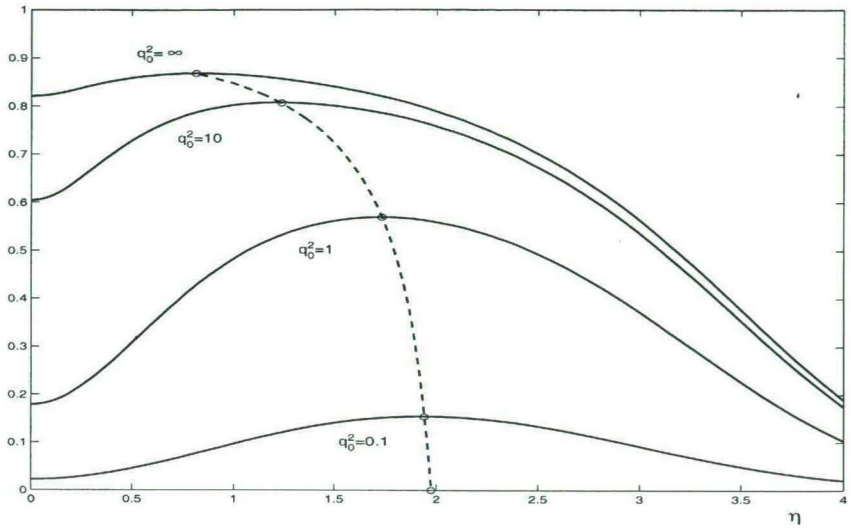

Figure 3. E(UR) and locus of max(E(UR)) ($m = 1$, $c = 1.96$).

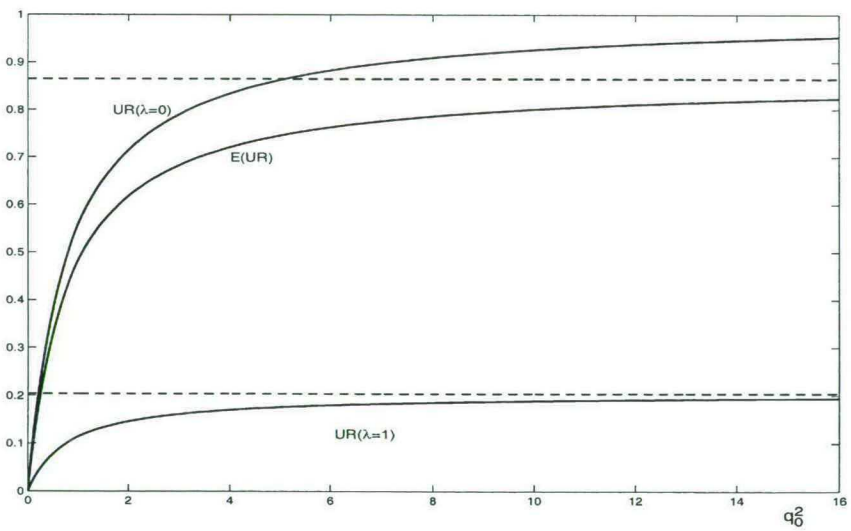

Figure 4. UR and E(UR) as a function of q_0^2 ($m = 1$, $c = 1.96$, $\eta = 1$).

We conclude that the effect of not reporting the true bias and variance of the pretest estimator can lead to serious misrepresentation of the results, *even in the case* $m = 1$. The larger is q_0^2 (known to the investigator), the larger will be the expected UR. For given q_0^2 we can draw the expected UR as a function of η, as in Figure 3, and calculate the maximum E(UR). Alternatively, we can calculate E(UR) at the point $\eta = \hat{\eta}$ and use this as an estimate of the seriousness of underreporting. The E(UR) can be as large as 0.87 (at $q_0^2 = \infty$ and $\eta = 0.82$). This means that in the worst case the expectation of the reported variance of the pretest estimator is only 13% of its actual mean squared error.

2.6 Model selection: general-to-specific and specific-to-general

When $m = 1$ pretesting is simple: look at the t-statistic for γ in the unrestricted model. If $|t| > c$, choose the unrestricted model (leading to b_u); otherwise choose the restricted model (leading to b_r). When $m > 1$ there are many ways to pretest. We consider the case $m = 2$ under the following conditions: model selection is based on t-statistics only, in the selected model all t-statistics are 'significant', and σ^2 is known.

Without loss of generality we normalize z_1 and z_2, the regressors associated with the nuisance parameters γ_1 and γ_2, by setting $z_i' M z_i = 1$ for $i = 1, 2$. Then,

$$Z'MZ = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

where $|r| < 1$, and

$$(Z'MZ)^{-1/2} = \frac{1}{\sqrt{1-r^2}} \begin{pmatrix} \alpha & -\rho \\ -\rho & \alpha \end{pmatrix},$$

with

$$\alpha = \frac{\sqrt{1+r} + \sqrt{1-r}}{2}, \qquad \rho = \frac{\sqrt{1+r} - \sqrt{1-r}}{2}.$$

There are four t-statistics to consider: two in the unrestricted model (denoted t_1 and t_2), one in the model where $\gamma_2 = 0$ (denoted $t_{(1)}$), and one in the model where $\gamma_1 = 0$ (denoted $t_{(2)}$). Let $\hat{\eta}_1$ and $\hat{\eta}_2$ denote the components of $\hat{\eta}$. Then,

each of the four t-statistics is a linear function of $\hat{\eta}_1$ and $\hat{\eta}_2$ in accordance with Theorem 1:

$$t_1 = \alpha\hat{\eta}_1 - \rho\hat{\eta}_2, \qquad t_2 = -\rho\hat{\eta}_1 + \alpha\hat{\eta}_2,$$

and

$$t_{(1)} = \alpha\hat{\eta}_1 + \rho\hat{\eta}_2, \qquad t_{(2)} = \rho\hat{\eta}_1 + \alpha\hat{\eta}_2.$$

Of course, since $\alpha^2+\rho^2 = 1$, all four t-statistics are normally distributed with unit variance and, under the appropriate null hypothesis, mean zero. Also, $t_{(1)}$ is independent of t_2 and $t_{(2)}$ is independent of t_1, for the same reason that b_r and $\hat{\eta}$ are independent. Further,

$$\text{corr}(t_1, t_{(1)}) = \text{corr}(t_2, t_{(2)}) = \sqrt{1-r^2} > 0,$$

and

$$\text{corr}(t_1, t_2) = -r, \qquad \text{corr}(t_{(1)}, t_{(2)}) = r.$$

Finally,

$$|t_1| > |t_2| \iff |t_{(1)}| > |t_{(2)}| \iff |\hat{\eta}_1| > |\hat{\eta}_2|.$$

A t-statistic is 'significant' if its absolute value exceeds some a priori chosen positive constant c, such as 1.96.

We shall investigate two pretest procedures that are in common use: 'general-to-specific' and 'specific-to-general'. Let $\mathcal{M}_0$ denote the restricted model, $\mathcal{M}_1$ the model with only z_1 ($\gamma_2 = 0$), $\mathcal{M}_2$ the model with only z_2 ($\gamma_1 = 0$), and $\mathcal{M}_{12}$ the unrestricted model. Then we define the general-to-specific (or 'backward' or 'top-down') procedure as follows:

a. Estimate the unrestricted model $\mathcal{M}_{12}$. This yields t-statistics t_1 and t_2;

b. Choose $\mathcal{M}_{12}$ if both t_1 and t_2 are significant;

c. Otherwise,

   (i) if $|t_1| > |t_2|$ estimate $\mathcal{M}_1$, yielding $t_{(1)}$. If $t_{(1)}$ is significant choose $\mathcal{M}_1$, otherwise choose $\mathcal{M}_0$;

   (ii) if $|t_1| \leq |t_2|$ estimate $\mathcal{M}_2$, yielding $t_{(2)}$. If $t_{(2)}$ is significant choose $\mathcal{M}_2$, otherwise choose $\mathcal{M}_0$.

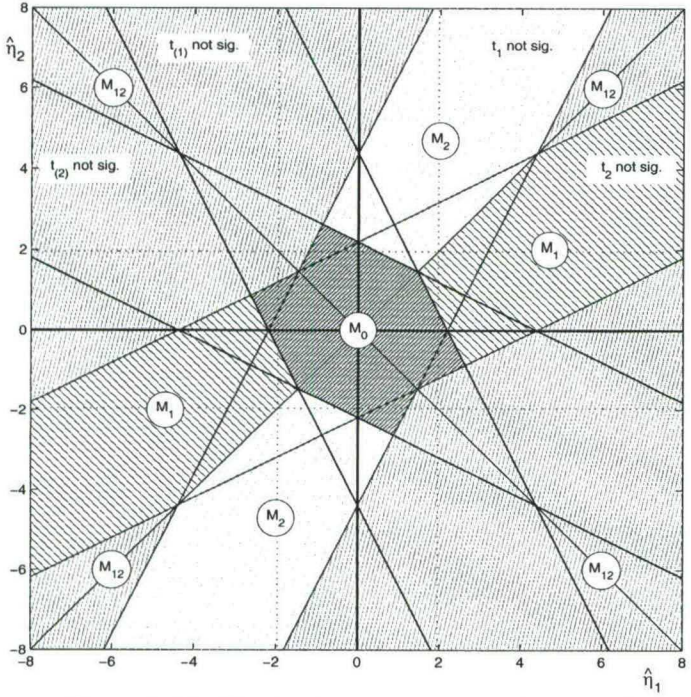

Figure 5. Model selection regions: general-to-specific.

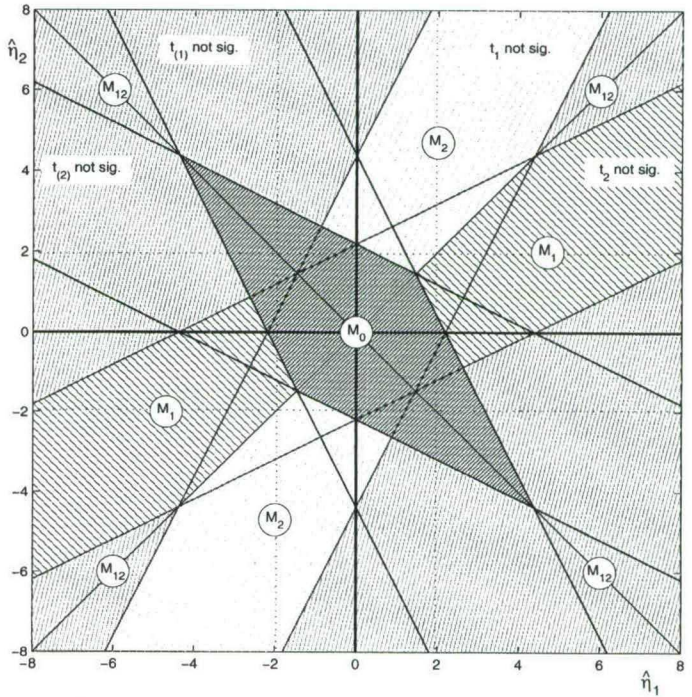

Figure 6. Model selection regions: specific-to-general.

Similarly, we define the specific-to-general (or 'forward' or 'bottom-up') procedure as follows:

a. Estimate both partially restricted models $\mathcal{M}_1$ and $\mathcal{M}_2$. This yields t-statistics $t_{(1)}$ and $t_{(2)}$;

b. Choose $\mathcal{M}_0$ if neither $t_{(1)}$ nor $t_{(2)}$ is significant;

c. Otherwise, estimate the unrestricted model yielding t_1 and t_2, and choose $\mathcal{M}_{12}$ if t_1 and t_2 are both significant;

d. In all other cases choose $\mathcal{M}_1$ (if $|t_{(1)}| > |t_{(2)}|$) or $\mathcal{M}_2$ (if $|t_{(1)}| \leq |t_{(2)}|$).

For $r = 0.8$, we graph the relevant regions in $(\hat{\eta}_1, \hat{\eta}_2)$-plane for both procedures in Figures 5 and 6.

Since the two cases ($|t_{(1)}| \leq c < |t_1|$, $|t_2| \leq c < |t_{(2)}|$) and ($|t_{(2)}| \leq c < |t_2|$, $|t_1| \leq c < |t_{(1)}|$) can not occur, we see that both procedures are identical, except for the case where t_1 and t_2 are both significant, while $t_{(1)}$ and $t_{(2)}$ are both not significant. In that case, the general-to-specific procedure chooses the unrestricted model and the specific-to-general procedure chooses the restricted model. In the special case $r = 0$, we find $t_1 = t_{(1)} = \hat{\eta}_1$ and $t_2 = t_{(2)} = \hat{\eta}_2$, and all pretest procedures coincide. When $|r| \to 1$, the difference between the two procedures is at its largest. In spite of the seemingly small difference between the two pretest procedures, the effect of pretesting on underreporting will be surprisingly different for the two procedures.

2.7 Underreporting with two nuisance parameters

In the case $m = 1$ the expected underreporting ratio E(UR) depends (for fixed c) on two parameters: q_0^2 (known to the investigator) and η (unknown). In the case $m = 2$, E(UR) depends, after normalization, on five parameters: q_0^2, q_1 and r (known), and η_1 and η_2 (unknown). In addition, E(UR) depends on the procedure.

We have four models to compare: the unrestricted $\mathcal{M}_{12}$, the partially restricted $\mathcal{M}_1$ ($\gamma_2 = 0$) and $\mathcal{M}_2$ ($\gamma_1 = 0$), and the restricted $\mathcal{M}_0$ ($\gamma_1 = \gamma_2 =$

0). This implies selection matrices $S_0 = I_2$, $S_1 = (0,1)'$, and $S_2 = (1,0)'$ (The matrix S_{12} has no columns), and hence $W_0 = 0$, $W_{12} = I_2$,

$$W_1 = \frac{1}{2}\begin{pmatrix} 1+\sqrt{1-r^2} & r \\ r & 1-\sqrt{1-r^2} \end{pmatrix},$$

and

$$W_2 = \frac{1}{2}\begin{pmatrix} 1-\sqrt{1-r^2} & r \\ r & 1+\sqrt{1-r^2} \end{pmatrix}.$$

Since $W = \lambda_0 W_0 + \lambda_1 W_1 + \lambda_2 W_2 + \lambda_{12} W_{12}$, we thus find

$$W = \frac{1}{2}\begin{pmatrix} \mathrm{tr}(W) + \sqrt{1-r^2}(\lambda_1 - \lambda_2) & r(\lambda_1+\lambda_2) \\ r(\lambda_1+\lambda_2) & \mathrm{tr}(W) - \sqrt{1-r^2}(\lambda_1-\lambda_2) \end{pmatrix},$$

where $\mathrm{tr}(W) = \lambda_1 + \lambda_2 + 2\lambda_{12}$. As before, let $\lambda(x) = 1$ if $|x| > c$ and 0 otherwise. Then,

$$\begin{aligned} \lambda_0 &= (1-\lambda(t_{(1)}))(1-\lambda(t_{(2)})) - \delta B_1, & \lambda_1 &= \lambda(t_{(1)})(1-\lambda(t_2)) - (1-\mu)B_2, \\ \lambda_2 &= \lambda(t_{(2)})(1-\lambda(t_1)) - \mu B_2, & \lambda_{12} &= \lambda(t_1)\lambda(t_2) - (1-\delta)B_1, \end{aligned}$$

with

$$\begin{aligned} B_1 &= \lambda(t_1)\lambda(t_2)(1-\lambda(t_{(1)}))(1-\lambda(t_{(2)})), \\ B_2 &= \lambda(t_{(1)})\lambda(t_{(2)})(1-\lambda(t_1))(1-\lambda(t_2)). \end{aligned}$$

Here, $\mu = 1$ if $|\hat{\eta}_1| > |\hat{\eta}_2|$ and 0 otherwise, and $\delta = 1$ if the pretest procedure is general-to-specific and 0 if the procedure is specific-to-general.

Because E(UR) depends on 5 parameters, only a 6-dimensional plot would do full justice to its behavior. This task being beyond us, let us first consider the mean squared error $R = \mathrm{MSE}(W\hat{\eta})$ and the expected reported variance $\mathrm{E}(W)$ for the two procedures. Both functions depend on η_1, η_2, and r. The $\mathrm{E}(W)$ is always bounded, as noted in Section 4. The matrix R is also bounded in the general-to-specific procedure, but R can be unbounded in the specific-to-general procedure. More specifically,

$$\max_{\eta_1,\eta_2} R(\eta_1,\eta_2,r) \to \infty \qquad \text{as } r \to 1,$$

when the procedure is specific-to-general. This very different behavior of R in the two procedures is reflected in Figure 7, where we consider

$$\mathrm{E}^{**}(\mathrm{UR}) = \max_{\eta_1,\eta_2} \mathrm{E}^{*}(\mathrm{UR}) = 1 - \min_{\eta_1,\eta_2} \min_{1\le j\le m} \xi_j(R^{-1/2}(\mathrm{E}\,W)R^{-1/2}),$$

as a function of r.[6]

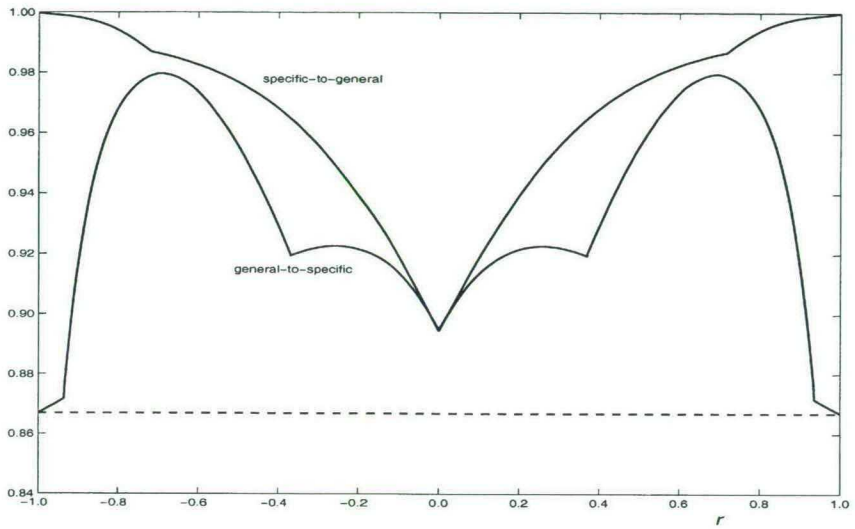

Figure 7. max(E(UR)) as a function of r ($m = 2$).

For both procedures the function E**(UR) is symmetric around $r = 0$. For $r = 0$ the two procedures are the same and the function value is almost 0.90. In the specific-to-general procedure, E**(UR) increases monotonically to 1 as r increases from 0 to 1. The general-to-specific procedure has a uniformly lower E**(UR), its behavior is non-monotonic, and it converges to 0.87 as $r \to 1$, the same maximum value as in the case $m = 1$ (depicted as a horizontal line in the figure). The difference between the two procedures is especially large when r is close to 1, that is when Mz_1 and Mz_2 are strongly correlated. This can be understood as follows. Let $r = 1$ and let $\eta_1 = -\eta_2 = \bar{\eta}$, say. Then, for large $\bar{\eta}$, the probability of choosing one of the partially restricted models $\mathcal{M}_1$ or $\mathcal{M}_2$ approaches 0. In the specific-to-general case, we will choose the restricted model $\mathcal{M}_0$ with probability approaching 0.95 and model $\mathcal{M}_{12}$ with probability approaching 0.05. Hence, for $r = 1$ and $\bar{\eta} \to \infty$, we find that E(UR) approaches 1 for any q_0^2. (In fact, the MSE of the pretest estimator is unbounded and proportional to $\bar{\eta}^2$ when $\bar{\eta}$ approaches

[6]The lines in Figures 7–11 are obtained by numerical techniques.

∞.) But in the general-to-specific case, the MSE is always bounded and hence $E^*(UR) < 1$, using (8).

Although the functions are continuous, there are various kinks. This is the result of the fact that there exist various local maxima. At a kink we move from one local maximum to another local maximum. Clearly, underreporting can be a very serious problem and, for $m \geq 2$, can be essentially unbounded, depending on the chosen pretest procedure.

For $r = 0$ the worst case gives $E^{**}(UR) = 0.87$ for $m = 1$ and 0.90 for $m = 2$. We now ask how underreporting depends on m. There are 2^m models to consider and one may think therefore that 'badness' increases by a factor of 2^m. On the other hand, all t-statistics are functions of only m random variables $\hat{\eta}_1, \ldots, \hat{\eta}_m$, so that 'badness' increases possibly only by a factor of m. We consider the special case where $Z'MZ = I_m$. Then all vectors Mz_i are orthogonal, and the m-dimensional problem collapses in to m one-dimensional problems (Theorem 3). All pretest procedures are the same in this case, and the maximum $E^{**}(UR)$ is plotted in Figure 8 as a function of m.

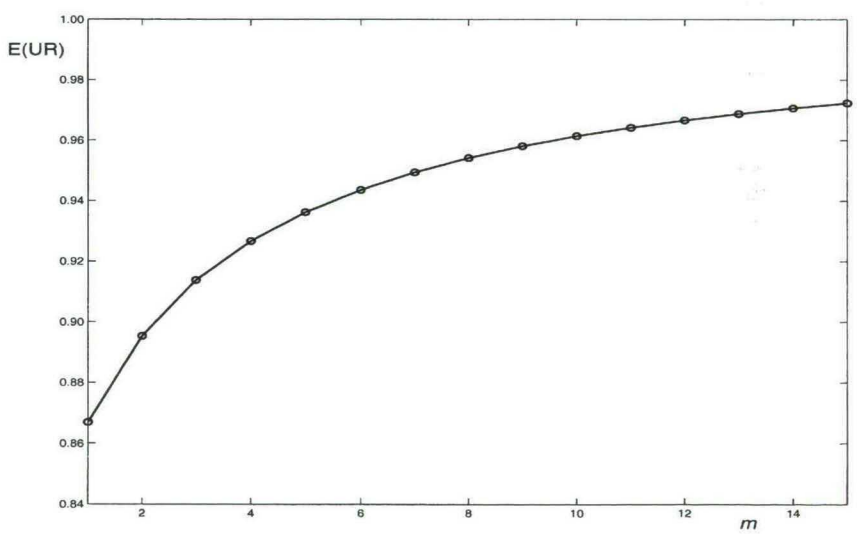

Figure 8. max(E(UR)) as a function of m ($Z'MZ = I_m$).

The figure reveals that $E^{**}(UR)$ increases with m but less than linearly. In

fact, we find that the actual pretest mean squared error is about $7.3m^{0.45}$ times the expected reported variance when $1 \leq m \leq 5$ and about $4.5m^{0.76}$ when $m > 6$. Although this result is valid only when $Z'MZ = I_m$, it nevertheless suggests that the increase in 'badness' is not as fast as one might have feared.

In a practical situation, we know q_0^2, q, and r, but not η_1 and η_2. Let us analyze one such situation where $q_0^2 = 2$, $q = (1/3, (2/3)\sqrt{2})'$ (so that $q'q = 1$), and $r = 0.8$.

Figures 9 and 10 give the E(UR) as a function of η_1 and η_2, first for the general-to-specific procedure, then for the specific-to-general procedure. The E(UR) lies always between 0 and 1, and is symmetric around the point $(\eta_1, \eta_2) = (0, 0)$. The functional dependence on (η_1, η_2) is quite complicated, and also quite different for the two procedures. In the general-to-specific procedure (Figure 9), E(UR) is 0 at $(\eta_1, \eta_2) = (4, -4)$, but can be as large as 0.6551 at $(0.4, 1.6)$. In the specific-to-general procedure (Figure 10), E(UR) varies from around 0 at $(4, 4)$ to 0.8798 around the point $(4, -4)$. In this case (and in general), the specific-to-general is more sensitive to underreporting than the general-to-specific procedure.

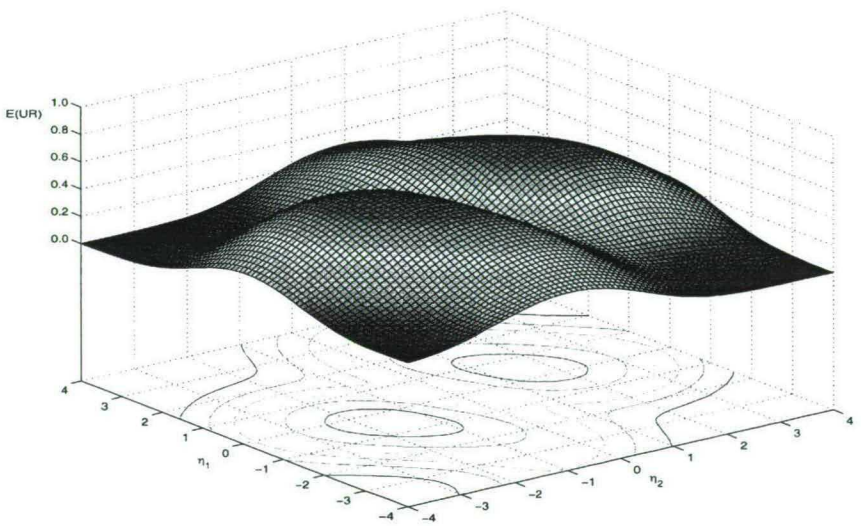

Figure 9. E(UR) as a function of η_1 and η_2: general-to-specific.

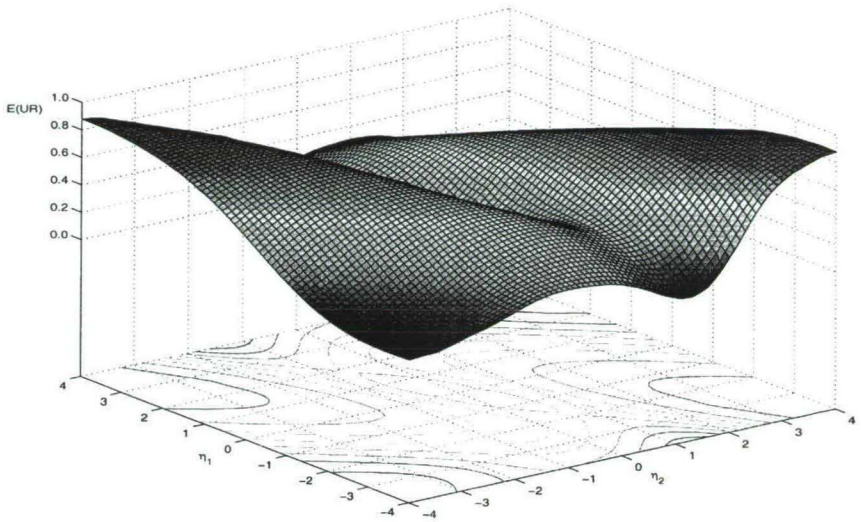

Figure 10. E(UR) as a function of η_1 and η_2: specific-to-general.

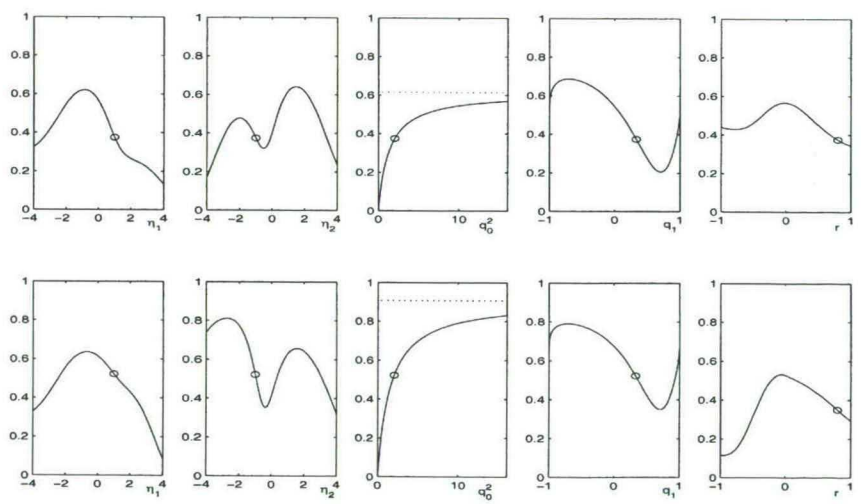

Figure 11. Sensitivity analysis for E(UR):
general-to-specific (top) and specific-to-general (bottom).

The contours in the (η_1, η_2) plane are iso-value curves: the darker (redder) the line, the higher the value.

Now consider a specific point $(\eta_1, \eta_2) = (1, -1)$. In Figure 11, we ask what happens in the 6-dimensional picture if we change the five parameters η_1, η_2, q_0^2, q_1, and r, one at a time.

At the chosen point, for both procedures, the E(UR) is an increasing function of q_0^2 (and q_2), but decreasing in η_1, η_2, q_1, and r. Figure 11 confirms that the E(UR) depends strongly, and not symmetrically, on η_1 and η_2. We already know that E(UR) is an increasing function of q_0^2, but the dependence is much less strong for the general-to-specific procedure than for the specific-to-general procedure. The E(UR) also depends strongly on q (that is q_1). Hence, different linear combinations of the β-parameters are affected differently by the pretest procedure. Sensitivity plots like Figure 11 can thus be used to assess the dependence of the E(UR) on the unknown parameters η_1 and η_2, and also on possible measurement error in the observed quantities q_0^2, q, and r.

2.8 Extensions and conclusions

In this paper we have analyzed the effect of ignoring the model selection procedure in reporting the bias and variance of the commonly used least-squares estimator. We conclude that underreporting is a very serious problem and that not reporting the correct pretest bias and variance can lead to very misleading results. The pretest bias appears to be less of a problem that the pretest variance.

When we have m auxiliary regressors $z_1, \ldots, z_m$, there are 2^m models to choose between. There are many different possible (viable) procedures to select the model. We find that the choice of model selection procedure (for example, general-to-specific or specific-to-general) matters a lot, and that the general-to-specific procedure seems to have more desirable properties. The influence of the selection procedure is higher when the correlation between the z variables (measured by $Z'MZ$) is high, than when it is low. If we can choose the auxiliary regressors such that they are 'orthogonal' (that is, $Z'MZ = I_m$), then all pretest procedures are the same, and hence the sampling properties of the estimators do not depend on the model selection procedure.

As the number of auxiliary regressors m grows, the dangers of underreporting grow as well, but less than linearly, in the sense that the MSE of the pretest estimator is approximately Am^{α} times the expected reported variance for some $0 < \alpha < 1$.

The paper shows not only that ignoring model selection can lead to serious underreporting, but also provides explicit formulae to calculate the correct bias, variance, and mean squared error, which are easy to implement in standard packages.

We now discuss briefly three extensions of the results obtained so far.

Unknown σ^2. Although Theorems 1 and 2 are valid whether or not σ^2 is known, the rest of the paper assumes that σ^2 is known. This is of course unrealistic and we need to address the question how the results are affected when σ^2 is unknown. As an example, let us consider the case of Figure 3 where $m = 1$, $q_0^2 = \infty$ and $c = 1.96$. When σ^2 is known, the E(UR) takes the values 0.82, 0.86, 0.79, and 0.19 for η equal to 0, 1, 2 and 4 respectively. When σ^2 is not known the calculations are more involved and depend on the degrees of freedom $n - k - m$. The results are summarized in Table 1.

	η			
$n-k-m$	0	1	2	4
10	0.76	0.83	0.77	0.26
30	0.80	0.85	0.78	0.22
50	0.81	0.86	0.79	0.21
∞	0.82	0.86	0.79	0.19

Table 1. E(UR) as a function of the d.f. $n - k - m$ (σ^2 unknown).

We see that the effects of estimating σ^2 are relatively small, especially in the region of interest where $|\eta|$ is around 1 or 2. Although this example is typical for the behavior of the E(UR), more work is needed in this direction, especially for $m \geq 2$.

Misspecification. We have also assumed that the unrestricted model is the data-generating process. Again, this may not be realistic, and we shall consider what happens if in fact a larger model generates the data. Thus, we assume that the data-generating process is $y = X\beta + Z_1\gamma_1 + Z_2\gamma_2 + \varepsilon$, but

Chapter 3

Forecast accuracy after pretesting with an application to the stock market*

3.1 Introduction

In econometrics we typically use the same data for both model selection and forecasting (and estimation). Standard statistical theory is therefore not directly applicable, because the properties of forecasts (and estimates) depend not only on the stochastic nature of the selected model, but also on the way this model was selected.

The simplest example of this situation is the standard linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon}$, where we are uncertain whether to include $\boldsymbol{z}$ or not. The usual procedure is to compute the t-statistic for γ, and then, depending on whether $|t|$ is 'large' or 'small', decide to use the unrestricted or the restricted (with $\gamma = 0$) model. We then forecast y_{n+1} from the selected model. This forecast is a *pretest* forecast, but we commonly report its properties as if forecasting had not been preceded by model selection. This is clearly wrong. We should correctly report the bias and variance (or mean squared error) of the forecasts, taking full account of the fact that model selection and forecasting are an integrated procedure. This paper attempts to do this, both in theory and practice.

*Joint paper with J.R. Magnus.

Section 2 contains the set-up and notation and reviews some earlier results, which are required for the development of the theory. The main result is presented in Section 3 (Theorem 1), giving the bias, variance, and mean squared forecast error of the pretest forecast (in fact, of the WALS forecast, a generalization of the pretest forecast). In Section 4 we apply the theory to the problem of forecasting stock market moves (Pesaran and Timmermann, 1994, 1995), and show that the recommendations of Pesaran and Timmermann are much less robust than naive econometrics would seem to imply, thus questioning the usefulness of the implied switching-portfolio strategy. In Section 5 we present a continuous analogue of pretesting which can greatly improve the properties of forecasts. In Section 6 we address the problem of how to incorporate the (obvious) fact that σ^2 is not known in our theory and applications. The effect of this extension is small. Some conclusions are offered in Section 7.

3.2 Set-up, notation, and preliminary results

The set-up is the same as in Magnus and Durbin (1999) and Danilov and Magnus (2001). We consider the standard linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{y}$ $(n \times 1)$ is the vector of observations, $\boldsymbol{X}$ $(n \times k)$ and $\boldsymbol{Z}$ $(n \times m)$ are matrices of nonrandom regressors, $\boldsymbol{\varepsilon}$ $(n \times 1)$ is a random vector of unobservable disturbances, and $\boldsymbol{\beta}$ $(k \times 1)$ and $\boldsymbol{\gamma}$ $(m \times 1)$ are unknown nonrandom parameter vectors.[1] We assume that $k \geq 1$, $m \geq 1$, $n-k-m \geq 1$, that the design matrix $(\boldsymbol{X} : \boldsymbol{Z})$ has full column-rank $k+m$, and that the disturbances $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$.[2]

The reason for distinguishing between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is that $\boldsymbol{X}$ contains explanatory variables ('focus' regressors) that we want in the model on theoretical or other grounds, while $\boldsymbol{Z}$ contains additional explanatory variables ('auxiliary' regressors) of which we are less certain.

[1] We follow the notation proposed in Abadir and Magnus (2002).

[2] In contrast to our estimation paper, we may allow $k = 0$ here, in which case $\boldsymbol{X}$ is absent. All subsequent results hold in that case, but some care needs to be taken about the interpretation of the formulas.

We define the matrices

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \quad \text{and} \quad \boldsymbol{Q} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2},$$

and the scaled parameter vector $\boldsymbol{\eta} = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{1/2}\boldsymbol{\gamma}/\sigma$. The least-squares (LS) estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $\boldsymbol{b}_u = \boldsymbol{b}_r - \boldsymbol{Q}\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{y}$, where $\boldsymbol{b}_r = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $\hat{\boldsymbol{\theta}} = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{1/2}\hat{\boldsymbol{\gamma}}$. The subscripts '$u$' and '$r$' denote 'unrestricted' and 'restricted' (with $\boldsymbol{\gamma} = \boldsymbol{0}$) respectively. Letting $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\theta}}/\sigma$, we see that $\hat{\boldsymbol{\eta}} \sim \mathrm{N}(\boldsymbol{\eta}, \boldsymbol{I}_m)$.

Let $\boldsymbol{S}_i$ be an $m \times r_i$ selection matrix of rank r_i $(0 \leq r_i \leq m)$, so that $\boldsymbol{S}_i' = (\boldsymbol{I}_{r_i} : \boldsymbol{O})$ or a column-permutation thereof. The equation $\boldsymbol{S}_i'\boldsymbol{\gamma} = \boldsymbol{0}$ thus selects a subset of the γ's to be equal to zero. Following Danilov and Magnus (2001), the LS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ under the restriction $\boldsymbol{S}_i'\boldsymbol{\gamma} = \boldsymbol{0}$ are then given by

$$\boldsymbol{b}_{(i)} = \boldsymbol{b}_r - \boldsymbol{Q}\boldsymbol{W}_i\hat{\boldsymbol{\theta}}, \qquad \boldsymbol{c}_{(i)} = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{W}_i\hat{\boldsymbol{\theta}},$$

where

$$\boldsymbol{W}_i = \boldsymbol{I}_m - \boldsymbol{P}_i, \qquad \boldsymbol{P}_i = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{S}_i\big(\boldsymbol{S}_i'(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1}\boldsymbol{S}_i\big)^{-1}\boldsymbol{S}_i'(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}$$

are symmetric idempotent $m \times m$ matrices of ranks $m - r_i$ and r_i respectively. (If $r_i = 0$ then $\boldsymbol{P}_i = \boldsymbol{O}$.) The distribution of $\boldsymbol{b}_{(i)}$ is given by

$$\boldsymbol{b}_{(i)} \sim \mathrm{N}\left(\boldsymbol{\beta} + \sigma\boldsymbol{Q}\boldsymbol{P}_i\boldsymbol{\eta},\ \sigma^2\left((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\boldsymbol{W}_i\boldsymbol{Q}'\right)\right). \tag{3.2}$$

There are 2^m different models to consider, one for each subset of $\gamma_1, \ldots, \gamma_m$ set equal to zero. A *pretest* estimator of $\boldsymbol{\beta}$ is obtained by first selecting one of these models (using t- or F-tests or other model selection criteria), and then estimating $\boldsymbol{\beta}$ in the selected model. We shall assume throughout that the model selection is based exclusively on the residuals from the unrestricted model, that is, on $\boldsymbol{M}\boldsymbol{y}$. This assumption appears to be satisfied in all standard cases. (Note that the residuals in the i-th model can always be expressed as $\boldsymbol{e}_{(i)} = \boldsymbol{D}_i\boldsymbol{M}\boldsymbol{y}$ for some idempotent matrix $\boldsymbol{D}_i$.) More generally, a WALS (weighted-average least-squares) estimator of $\boldsymbol{\beta}$ is defined as $\boldsymbol{b} = \sum_i \lambda_i \boldsymbol{b}_{(i)}$, where the weights satisfy $\lambda_i = \lambda_i(\boldsymbol{M}\boldsymbol{y})$, $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$, and the sum is taken over all 2^m models. Clearly, the pretest estimator is a special case of the WALS estimator when all λ_i's are 0 except one which is 1.

The WALS estimator can be written as $\boldsymbol{b} = \boldsymbol{b}_r - \boldsymbol{Q}\boldsymbol{W}\hat{\boldsymbol{\theta}}$, where $\boldsymbol{W} = \boldsymbol{I}_m - \boldsymbol{P}$ and $\boldsymbol{P} = \sum_i \lambda_i \boldsymbol{P}_i$. (Notice that both $\boldsymbol{P}$ and $\boldsymbol{W}$ are random matrices,

because the $\{\lambda_i\}$ are random.) The equivalence theorem (for estimation) now says that

$$\mathrm{E}(\boldsymbol{b}) = \boldsymbol{\beta} - \sigma \boldsymbol{Q}\,\mathrm{E}(\boldsymbol{W}\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}), \quad \mathrm{var}(\boldsymbol{b}) = \sigma^2\left((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\,\mathrm{var}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{Q}'\right),$$

and hence that

$$\mathrm{MSE}(\boldsymbol{b}) = \sigma^2\left((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\,\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{Q}'\right),$$

showing that the properties of the complicated WALS (pretest) estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ depend critically on the properties of the less complicated estimator $\boldsymbol{W}\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$.

3.3 The equivalence theorem for forecasting

Suppose now that our interest is in forecasting rather than estimation. We assume that the data are generated by (3.1), possibly with one or more of the γ_i equal to zero. Under the restriction $\boldsymbol{S}_i'\boldsymbol{\gamma} = \boldsymbol{0}$ the one-period-ahead LS forecast is given by

$$\begin{aligned}\hat{y}_{n+1}^{(i)} &= \boldsymbol{x}_{n+1}'\boldsymbol{b}_{(i)} + \boldsymbol{z}_{n+1}'\boldsymbol{c}_{(i)} \\ &= \boldsymbol{x}_{n+1}'\left(\boldsymbol{b}_r - \boldsymbol{Q}\boldsymbol{W}_i\hat{\boldsymbol{\theta}}\right) + \boldsymbol{z}_{n+1}'\left((\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{W}_i\hat{\boldsymbol{\theta}}\right) \\ &= \boldsymbol{x}_{n+1}'\boldsymbol{b}_r - \boldsymbol{\omega}'\boldsymbol{W}_i\hat{\boldsymbol{\theta}} = \boldsymbol{x}_{n+1}'\boldsymbol{b}_r - \sigma\boldsymbol{\omega}'\boldsymbol{W}_i\hat{\boldsymbol{\eta}},\end{aligned}$$

where

$$\boldsymbol{\omega} = \boldsymbol{Q}'\boldsymbol{x}_{n+1} - (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{z}_{n+1},$$

and $\boldsymbol{x}_{n+1}$ and $\boldsymbol{z}_{n+1}$ denote next period's values of the focus and auxiliary regressors respectively. Since the actual choice of model is uncertain and depends on the data and the model selection procedure, the forecast could be based on any of the 2^m available models (or a linear combination thereof). Hence the WALS forecast takes the form

$$\hat{y}_{n+1} = \sum_i \lambda_i \hat{y}_{n+1}^{(i)} = \boldsymbol{x}_{n+1}'\boldsymbol{b}_r - \sigma\boldsymbol{\omega}'\boldsymbol{W}\hat{\boldsymbol{\eta}}. \tag{3.3}$$

Notice that $\sigma\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\theta}}$ and can thus be observed, but that nevertheless $\hat{y}_{n+1}$ depends on σ, because $\boldsymbol{W}$ (through λ_i) depends on σ.

Since $y_{n+1} = \boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma} + \varepsilon_{n+1}$, we obtain the forecast error (FE) as

$$\begin{aligned}\text{FE} &= \hat{y}_{n+1} - y_{n+1} \\ &= \boldsymbol{x}'_{n+1}(\boldsymbol{b}_r - \boldsymbol{\beta}) - \sigma\boldsymbol{\omega}'\boldsymbol{W}\hat{\boldsymbol{\eta}} - \sigma\boldsymbol{z}'_{n+1}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{\eta} - \varepsilon_{n+1} \\ &= \boldsymbol{x}'_{n+1}(\boldsymbol{b}_r - \boldsymbol{\beta} - \sigma\boldsymbol{Q}\boldsymbol{\eta}) - \sigma\boldsymbol{\omega}'(\boldsymbol{W}\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) - \varepsilon_{n+1}.\end{aligned}$$

The following properties of the forecast error can now be established.

Theorem 1 (Equivalence theorem for forecasting): The WALS forecast error has the following expectation, variance, and mean squared error:

$$\mathrm{E}(\text{FE}) = -\sigma\boldsymbol{\omega}'\,\mathrm{E}(\boldsymbol{W}\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}),$$

$$\mathrm{var}(\text{FE}) = \sigma^2\left(\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \boldsymbol{\omega}'\,\mathrm{var}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{\omega} + 1\right),$$

and hence

$$\text{MSFE} = \sigma^2\left(\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \boldsymbol{\omega}'\,\text{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{\omega} + 1\right).$$

Proof: The essential ingredient is that $\boldsymbol{b}_r$ and $\boldsymbol{M}\boldsymbol{y}$ are independent, because they are jointly normal and uncorrelated since $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{O}$. This implies that $\boldsymbol{b}_r$ and $\boldsymbol{W}\hat{\boldsymbol{\eta}}$ are independent, and hence that $(\boldsymbol{b}_r, \boldsymbol{W}\hat{\boldsymbol{\eta}}, \varepsilon_{n+1})$ are all independent of each other. The results follow. ||

The importance of Theorem 1 is twofold. First, it gives explicit expressions for the first two moments of the forecast error, where we notice that these moments depend on $\boldsymbol{\eta}$ and σ^2, but not on $\boldsymbol{\beta}$. Secondly, it helps us to find an *optimal* forecast. If we can find λ_i's such that $\boldsymbol{W}\hat{\boldsymbol{\eta}}$ is an optimal estimator of $\boldsymbol{\eta}$ (in the sense of minimizing the mean squared error), then *the same* λ_i's will provide an optimal forecast. (These λ_i's are also the ones which provide the optimal WALS estimator of $\boldsymbol{\beta}$.) The question of finding an optimal estimator of $\boldsymbol{\eta}$ was studied in Magnus (2002), and led to the 'neutral Laplace' estimator. In Section 5 we shall apply the Laplace weights to forecasting, and demonstrate the superiority of this approach.

Theorem 1 thus gives the actual (true) moments of the forecast error, taking into account that pretesting has occurred. In a typical applied paper, however, one does not take pretesting into account. Consequently, the bias of the forecast (that is, the expectation of the forecast error) is reported to be zero, the reported MSFE (variance), denoted $\widetilde{\text{MSFE}}$, is given by

$$\widetilde{\text{MSFE}} = \sigma^2\left(\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \boldsymbol{\omega}'\boldsymbol{W}\boldsymbol{\omega} + 1\right), \tag{3.4}$$

and the reported 95% prediction interval for y_{n+1} is

$$\hat{y}_{n+1} \pm 1.96\sigma\sqrt{\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \boldsymbol{\omega}'\boldsymbol{W}\boldsymbol{\omega} + 1}, \tag{3.5}$$

where σ is estimated by some consistent estimator $\hat{\sigma}$. In contrast, if we take proper account of the effects of model selection, then the actual value of the forecast $\hat{y}_{n+1}$ remains the same, but its moments are quite different. Let us define the two functions

$$\psi_1(\boldsymbol{\eta}) := \boldsymbol{\omega}' \,\mathrm{E}(\boldsymbol{W}\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \tag{3.6}$$

and

$$\psi_2(\boldsymbol{\eta}) := \boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \boldsymbol{\omega}' \,\mathrm{var}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{\omega}, \tag{3.7}$$

both of which depend also on σ, because $\boldsymbol{W}$ depends on σ. Then, by Theorem 1,

$$\mathrm{FE} \sim \left(-\sigma\psi_1(\boldsymbol{\eta}),\ \sigma^2\left(\psi_2(\boldsymbol{\eta}) + 1\right)\right),$$

so that an approximate 95% prediction interval for y_{n+1} is given by

$$\hat{y}_{n+1} + \sigma\left(\psi_1(\boldsymbol{\eta}) \pm 1.96\sqrt{\psi_2(\boldsymbol{\eta}) + 1}\right). \tag{3.8}$$

The interval is approximate because the distribution of FE is not normal. Furthermore, in contrast to (3.5), the interval depends on $\boldsymbol{\eta}$ (and on σ of course), which is unknown. We obtain an estimated prediction interval by replacing $\boldsymbol{\eta}$ and σ by the estimates $\hat{\boldsymbol{\eta}}$ and $\hat{\sigma}$.

When the number of observations n becomes large, then $\hat{\sigma}$ will converge to σ, but $\hat{\boldsymbol{\eta}}$ will not converge to $\boldsymbol{\eta}$, because $\mathrm{var}(\hat{\boldsymbol{\eta}}) = \boldsymbol{I}_m$. Hence, $\hat{\boldsymbol{\eta}}$ is an unbiased but not a consistent estimator of $\boldsymbol{\eta}$. To protect ourselves against 'large' deviations of $\hat{\boldsymbol{\eta}}$ from $\boldsymbol{\eta}$, we shall also consider the more conservative interval

$$\hat{y}_{n+1} + \sigma C_1(\hat{\boldsymbol{\eta}}) < y_{n+1} < \hat{y}_{n+1} + \sigma C_2(\hat{\boldsymbol{\eta}}), \tag{3.9}$$

where

$$C_1(\hat{\boldsymbol{\eta}}) := \min_{\boldsymbol{\eta}\in\mathcal{H}(\hat{\boldsymbol{\eta}})}\left(\psi_1(\boldsymbol{\eta}) - 1.96\sqrt{\psi_2(\boldsymbol{\eta}) + 1}\right)$$

and

$$C_2(\hat{\boldsymbol{\eta}}) := \max_{\boldsymbol{\eta}\in\mathcal{H}(\hat{\boldsymbol{\eta}})}\left(\psi_1(\boldsymbol{\eta}) + 1.96\sqrt{\psi_2(\boldsymbol{\eta}) + 1}\right).$$

The set $\mathcal{H}$ is an m-dimensional cube, defined by $\mathcal{H}(\hat{\boldsymbol{\eta}}) := \{\boldsymbol{\eta} : |\hat{\eta}_i - \eta_i| < a_m,\ i = 1, \ldots, m\}$, where a_m is determined such that, for standard-normal u, $\Pr(|u| < a_m)^m = 0.95$. (In our application, $m = 4$ and hence $a_m = 2.49$.)

3.4 Forecasting stock returns

In order to investigate the effects of ignoring pretesting on forecasts in practice, we will consider a question from the finance literature. Perhaps the first important application of linear regression in finance is the capital asset pricing model. Black, Jensen and Scholes (1972) proposed a linear regression model to explain empirically observed asset returns. Fama and MacBeth (1973) introduced a cross-section approach, and regressed the asset's excess return on the intercept and the β's of the CAPM model. Subsequent studies extended the set of explanatory variables. Equity risk premia related variables, such as the dividend yield, were suggested by Rozeff (1984), while French, Schwert and Stambaugh (1987) proposed default bond premia. Fama and French (1989) suggested to use the interest rate as an explanatory variable, since it affects the overall economic activity and, as a consequence, the stock market activity. Using the inflation rate (or an inflation-related characteristic) as an explanatory variable goes back to Lucas (1976). Industrial production variables were used by Chen, Roll and Ross (1986) and Balvers, Cosimano and McDonald (1990). A price-earnings variable, describing the relationship of the stock price and the actual earnings of the company, was used in Fama and French (1992). Inspired by the development of regression models, Cheng, Lo and Ma (1990) attempted to forecast the Hong Kong stock price index. Their regression models were however not sufficiently powerful to effectively predict the direction of the change in the index. Pesaran and Timmermann (1994) were more successful and demonstrated that a regression model preceded by model selection can actually predict movements of the Dow Jones and Standard & Poor 500 indexes with a sufficient degree of accuracy. This result was enriched and reinforced in Pesaran and Timmermann (1995), where a number of model selection criteria were employed. The problem of forecasting the market moves was reconsidered in Granger and Pesaran (2000), where the authors argue that not a point stock value but rather the probability of the fall in the stock market is the key element, and propose a way to estimate this probability.

We shall reconsider the question discussed by Pesaran and Timmermann (1994), hereafter PT94: can the annual excess returns on common stocks for the Standard & Poor 500 (SP 500) index be predicted?[3] Of course, PT94

[3]PT94 also consider the Dow Jones Industrial portfolio, and also monthly and quarterly frequencies. We shall only consider the SP 500 index and annual returns.

pretested. In fact, they state explicitly (p. 339) that they "experimented with a number of specifications". The dependent variable in the linear regression is ρ_t, the excess returns in year t. In analyzing the effect of pretesting we have to decide which regressors play a role and which of these are focus regressors and which are auxiliary. The distinction is not completely unambiguous, but we decided — after carefully studying their model selection description — that PT94's model contains four focus regressors ($k = 4$) and four auxiliary regressors ($m = 4$).[4] The focus regressors are:

constant term,
PI_{t-2}: annual inflation rate (lagged two periods),
DI3_{t-1}: change in 3-month T-bill rate (lagged one period),
TERM_{t-1}: term premium (lagged one period),

and the auxiliary regressors are:

YSP_{t-1}: dividend yield on SP 500 portfolio (lagged one period),
DIP_{t-1}: annual change in industrial production (lagged one period),
PER_{t-1}: price-earnings ratio (lagged one period),
DLEAD_{t-2}: annual change in leading business cycle indicator (lagged two periods).

Employing a forward (specific-to-general) model selection procedure, PT94 (p. 339) then obtain the following estimated model of the annual excess returns over the period 1954–1991:

$$\begin{array}{lllll} \hat{\rho}_t = -0.289 - & 1.72\,\text{PI}_{t-2} - & 0.06\,\text{DI3}_{t-1} + & 0.11\,\text{TERM}_{t-1} + & 9.17\,\text{YSP}_{t-1}. \\ \quad (0.077) & \quad (0.44) & \quad (0.02) & \quad (0.04) & \quad (2.02) \end{array}$$

We could not acquire *exactly* the same data set as PT94, but we almost could. In addition, since our data set extends to the year 2001, we had to employ a slightly different definition of the term premium TERM_{t-1}.[5] Our data set thus contains eight annual time series (plus a constant term) over

[4] In fact, PT94 did more pretesting than we analyze in this paper, so that $m > 4$ and the effect of ignoring pretesting is even larger than we report.

[5] PT94 measure the term premium as the difference between the 6-month commercial paper rate (risky) and the 3-month T-bill rate (riskless) in January. Since the 6-month commercial paper rate does not exist after 1997, we use the 3-month financial paper rate instead.

46 years (1956–2001).[6] A full description and all the data are given in the appendix.

With our data set we re-estimated the annual excess returns over the period 1956–1991, also employing a forward pretest procedure. This led to the same model as obtained by PT94, but to slightly different estimates:

$$\hat{\rho}_t = \underset{(0.084)}{-0.343} - \underset{(0.44)}{1.65\,\mathrm{PI}_{t-2}} - \underset{(0.02)}{0.04\,\mathrm{DI3}_{t-1}} + \underset{(0.04)}{0.17\,\mathrm{TERM}_{t-1}} + \underset{(2.17)}{10.14\,\mathrm{YSP}_{t-1}}.$$

A few words of explanation are in order. First, the *forward* pretest procedure (also called specific-to-general) is defined by starting from the smallest model (the restricted model) with k explanatory variables (the $\boldsymbol{X}$-variables). We first estimate the m models with one additional regressor. If none of the m t-statistics is significant, we choose the restricted model. If at least one of the t-statistics is significant, we select the regressor whose t-statistic is the largest (in absolute value), and keep this regressor in the model, whatever happens later in the procedure. Next, we estimate the $m-1$ models with two additional regressors, one of which is the one already selected. Proceeding in this way, we always select a model in an well-defined and unambiguous manner. Notice however that in the final model there is no guarantee that all t-statistics are significant.

Secondly, the t-statistics are computed in the traditional manner, that is, using an estimate of σ^2 based on the submodel under consideration. In this way, we mimic precisely what happens in applied work. The critical value, however, is always takes to be 1.96. This does not make any serious difference, and is more in line with the normality assumptions made in the approximations.

Thirdly, all computations are performed by Monte-Carlo methods, based on 1,000 replications, and properly tested for stability.

We now discuss the effect of pretesting on the forecasts. The forecasts discussed below are one-period-ahead forecasts for the period 1992–2001, based on all information available at the moment of forecasting. For example, the forecast for the year 2000 is based on the model selected and estimated using the 1956–1999 data. It is thus possible (and indeed it happens) that the forecast in one year is based on a different model than in another year.

[6] We could not obtain the full data set for 1954 and 1955, because TERM and YSP are not available in 1953 and 1954.

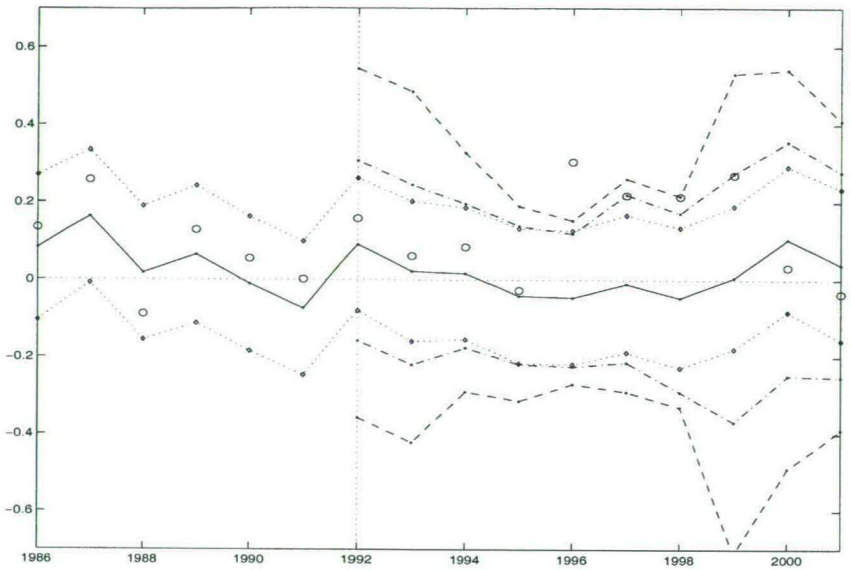

Figure 1. Pretest forecasts $\hat{y}_{n+1}$ with three sets of prediction intervals.

In Figure 1, the solid line gives the one-period-ahead forecasts $\hat{y}_{n+1}$, while the little open circles give the realized values y_{n+1}. The forecasts are the same, whether we take pretesting into account or not. The difference lies in the prediction bounds. The two dotted lines give the standard least-squares 95% prediction bounds (ignoring the effects of pretesting) as given in (3.5). These are the prediction bounds as would have been reported by PS94. They are symmetric around $\hat{y}_{n+1}$. We see that only 60% of the forecasts (six out of ten) lie in this standard prediction interval.

The two dash-dotted lines show the approximate 95% prediction bounds of the pretest forecast, based on (3.8), while the dashed lines give the more conservative interval, based on (3.9). Because of the bias effect, these intervals are not symmetric around $\hat{y}_{n+1}$. Now 80% of the forecasts lie in the approximate 95% prediction interval, and 90% in the more conservative interval. The year 1996 appears to be the most difficult to predict, partly because the market changed direction between 1995 and 1996.

In doing the calculations for the dash-dotted and the dashed intervals, we

estimate σ^2 by the LS estimator in the unrestricted model, that is,

$$\hat{\sigma}^2 = \frac{1}{n-k-m}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_u - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_u - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}), \tag{3.10}$$

which simplifies the calculations without effecting the results; see Section 6.[7]

Although the WALS forecast is seriously biased in some years, and the standard deviation is seriously underestimated, and therefore standard prediction intervals can be very misleading for evaluating the accuracy of the forecast, the difference between the dotted and the dash-dotted lines is not spectacularly large, on average only 1.3 times as wide. Hence, ignoring the effects of pretesting on the prediction bounds of the forecast is not necessarily disastrous, at least within the restrictions of the PS94 data set.

Lack of sensitivity in one direction does not, however, imply lack of sensitivity in another direction. For the question posed in PS94, the most important estimate is not the forecast, but rather — as argued by Granger and Pesaran (2000) — the forecast probability $\Pr(y_{n+1} > 0)$. Here the effect of ignoring pretesting will turn out to be rather more dramatic.

Since the error term is assumed to be normally distributed, we have

$$\begin{aligned}\Pr(y_{n+1} > 0) &= \Pr\left(\boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma} + \varepsilon_{n+1} > 0\right) \\ &= \Pr\left(-\varepsilon_{n+1} < \boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma}\right) \\ &= \Phi\left(\frac{\boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma}}{\sigma}\right),\end{aligned} \tag{3.11}$$

where $\Phi(\cdot)$ denotes the standard-normal c.d.f. If the value of $\Pr(y_{n+1} > 0)$ is larger than 0.5, the investor will conclude that the market will go up in the next period, and therefore will invest in stocks, if risk neutrality is assumed. If, on the other hand, the value of $\Pr(y_{n+1} > 0)$ is smaller than 0.5, the investor will conclude that the market will go down and will invest in bonds. Of course, the probability that $\Pr(y_{n+1} > 0)$ is not known and needs to be estimated by $\Phi(\hat{y}_{n+1}/\hat{\sigma})$. Moreover, we want to know how good the estimates are, using appropriate prediction intervals.

[7]Notice that ψ_1 and ψ_2, and therefore C_1 and C_2, also depend on σ.

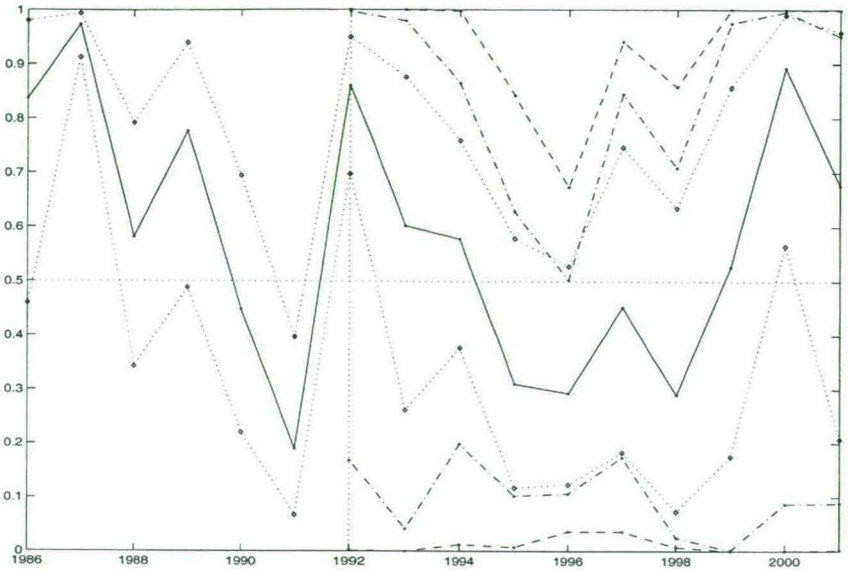

Figure 2. Pretest forecast probabilities $\Pr(y_{n+1} > 0)$ with three sets of prediction intervals.

The solid line in Figure 2 gives the estimated probability that $\Pr(y_{n+1} > 0)$. If we take no account of the effects of pretesting, then a 95% prediction interval for the parameter $(\boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma})/\sigma$ is given by

$$\frac{\hat{y}_{n+1}}{\sigma} \pm 1.96\sqrt{\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \boldsymbol{\omega}'\boldsymbol{W}\boldsymbol{\omega}}, \tag{3.12}$$

and the dotted lines are based on these bounds.

If, however, we do take account of pretesting, then

$$\hat{y}_{n+1} \sim \left(\boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma} - \sigma\psi_1(\boldsymbol{\eta}),\ \sigma^2\psi_2(\boldsymbol{\eta})\right), \tag{3.13}$$

where ψ_1 and ψ_2 are defined in (3.6) and (3.7), so that an approximate 95% prediction interval for $(\boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma})/\sigma$ is given by

$$\frac{\hat{y}_{n+1}}{\sigma} + \psi_1(\boldsymbol{\eta}) \pm 1.96\sqrt{\psi_2(\boldsymbol{\eta})}. \tag{3.14}$$

This interval depends on $\boldsymbol{\eta}$ (and σ) which is unknown. We obtain an estimated prediction interval by replacing $\boldsymbol{\eta}$ and σ by the estimates $\hat{\boldsymbol{\eta}}$ and $\hat{\sigma}$, leading to the dash-dotted lines.

Finally, as in Section 3, we obtain more conservative bounds, taking into account that $\hat{\boldsymbol{\eta}}$, although unbiased, is inconsistent:

$$\frac{\hat{y}_{n+1}}{\sigma} + C_3(\hat{\boldsymbol{\eta}}) < \frac{\boldsymbol{x}'_{n+1}\boldsymbol{\beta} + \boldsymbol{z}'_{n+1}\boldsymbol{\gamma}}{\sigma} < \frac{\hat{y}_{n+1}}{\sigma} + C_4(\hat{\boldsymbol{\eta}}), \tag{3.15}$$

where

$$C_3(\hat{\boldsymbol{\eta}}) := \min_{\boldsymbol{\eta}\in\mathcal{H}(\hat{\boldsymbol{\eta}})} \left(\psi_1(\boldsymbol{\eta}) - 1.96\sqrt{\psi_2(\boldsymbol{\eta})}\right),$$

and

$$C_4(\hat{\boldsymbol{\eta}}) := \max_{\boldsymbol{\eta}\in\mathcal{H}(\hat{\boldsymbol{\eta}})} \left(\psi_1(\boldsymbol{\eta}) + 1.96\sqrt{\psi_2(\boldsymbol{\eta})}\right).$$

While the standard regression prediction intervals are already large, allowing only two years (1992, 2000) where a direction can be forecasted with any confidence, the (correct) pretest prediction intervals are such that we cannot be confident in *any* year. This is true for the dash-dotted lines and, *a forteriori*, for the more conservative dashed lines.

The difference between the dotted and the dash-dotted lines is twice as large as in Figure 1, on average 2.6 times as wide. This large difference in the effects of ignoring pretesting between Figures 1 and 2 (the dash-dotted lines) can be attributed completely to a small difference between formulas (3.8) and (3.14). The first formula contains the term $\psi(\boldsymbol{\eta}) + 1$, which is replaced by $\psi(\boldsymbol{\eta})$ in the second formula. The simple appearance of one 1 thus appears to have a large effect on the bounds.

We conclude that ignoring the effects of pretesting on the distribution of the forecast can lead to a serious misrepresentation. The pretest forecast is biased and has a larger variance than is apparent from the regression results. The one-period-ahead forecasts are much less precise than naive econometrics would lead us to believe. The effects of pretesting of forecasting are thus serious and should be analyzed and incorporated in econometric analyses.

3.5 Optimal forecasts using the Laplace weights

We have seen that in evaluating the properties of forecasts, especially forecast probabilities, we need to take the model selection aspect into account. So

far, we have only considered the standard pretest procedure, where we first select the 'best' model and then forecast on the basis of this selected model. Such a procedure is discontinuous and hence inadmissible. Since we are not in the business of finding the 'best' model, but rather of finding the 'best' forecast, we may wish to consider a (continuous) weighted average of models instead of the (discontinuous) pretest model selection. But which weights should be taken? In Magnus (2002) a Bayesian solution to this problem is proposed (in the estimation context) for the case $m = 1$. When $m = 1$, there are only two possible models, the restricted (r) and the unrestricted (u), and the forecast takes the simple form (see (3.3))

$$\hat{y}_{n+1} = \lambda \hat{y}_{n+1}^{(u)} + (1 - \lambda)\hat{y}_{n+1}^{(r)}.$$

The proposed weight-function $\lambda = \lambda(\hat{\eta})$ is

$$\lambda(\hat{\eta}) = \frac{\int \eta \pi(\eta) \exp(-(\hat{\eta} - \eta)^2/2)\, \mathrm{d}\eta}{\hat{\eta} \int \pi(\eta) \exp(-(\hat{\eta} - \eta)^2/2)\, \mathrm{d}\eta},$$

where the prior π is the 'neutral' Laplace density,

$$\pi(\eta) = \frac{c}{2} \exp\left(-c\,|\eta|\right), \quad -\infty < \eta < \infty, \quad c = \log 2.$$

The neutrality of the prior guarantees that $\text{median}(\eta) = 0$ and $\text{median}(\eta^2) = 1$. We know that the use of the Laplace weights leads to better estimates (admissible to begin with) than the pretest weights.

When $m > 1$ it is not so clear how the weights should be taken. However, in the special case where $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$, the multi-dimensional problem separates into m one-dimensional problems, and we can use the Laplace weights for each dimension separately; see Danilov and Magnus (2001, Theorem 3).

Let us consider the 'orthogonalization' $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$ in some more detail. Orthogonalization can always be achieved by taking appropriate linear combinations of the m auxiliary regressors in $\boldsymbol{Z}$ (leaving the focus regressors unchanged). More specifically, let $\boldsymbol{T}_1$ be an orthogonal $m \times m$ matrix such that $\boldsymbol{T}_1'\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}\boldsymbol{T}_1 = \boldsymbol{\Lambda}$ (diagonal). Then, letting $\boldsymbol{T} = \boldsymbol{T}_1\boldsymbol{\Lambda}^{-1/2}$, we have $\boldsymbol{T}'\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}\boldsymbol{T} = \boldsymbol{I}_m$. Now define new auxiliary regressors $\boldsymbol{Z}^* = \boldsymbol{Z}\boldsymbol{T}$ and $\boldsymbol{z}_{n+1}^* = \boldsymbol{T}'\boldsymbol{z}_{n+1}$. Then, clearly, $\boldsymbol{Z}^{*\prime}\boldsymbol{M}\boldsymbol{Z}^* = \boldsymbol{I}_m$. As a consequence of this transformation, $\boldsymbol{\omega}$, $\boldsymbol{R}(\boldsymbol{\eta}) := \text{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})$, and MSFE will all change, but $\boldsymbol{\omega}'\boldsymbol{\omega}$ will not change. This follows because

$$\begin{aligned}\boldsymbol{\omega} &= \boldsymbol{Q}'\boldsymbol{x}_{n+1} - (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{z}_{n+1} \\ &= (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\left(\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} - \boldsymbol{z}_{n+1}\right),\end{aligned}$$

so that

$$\begin{aligned}\boldsymbol{\omega}^* &= (\boldsymbol{Z}^{*\prime}\boldsymbol{M}\boldsymbol{Z}^*)^{-1/2}\left(\boldsymbol{Z}^{*\prime}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} - \boldsymbol{z}^*_{n+1}\right)\\ &= \boldsymbol{T}'\left(\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} - \boldsymbol{z}_{n+1}\right).\end{aligned}$$

Then the fact that $\boldsymbol{T}\boldsymbol{T}' = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1}$ implies that $\boldsymbol{\omega}^{*\prime}\boldsymbol{\omega}^* = \boldsymbol{\omega}'\boldsymbol{\omega}$. The only difference between

$$\begin{aligned}\text{MSFE} &= \sigma^2\left(\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \boldsymbol{\omega}'\boldsymbol{R}(\boldsymbol{\eta})\boldsymbol{\omega} + 1\right)\\ &= \sigma^2(\boldsymbol{\omega}'\boldsymbol{\omega})\left(\frac{\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}+1}{\boldsymbol{\omega}'\boldsymbol{\omega}} + \frac{\boldsymbol{\omega}'\boldsymbol{R}(\boldsymbol{\eta})\boldsymbol{\omega}}{\boldsymbol{\omega}'\boldsymbol{\omega}}\right)\end{aligned}$$

and

$$\text{MSFE}^* = \sigma^2(\boldsymbol{\omega}^{*\prime}\boldsymbol{\omega}^*)\left(\frac{\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}+1}{\boldsymbol{\omega}^{*\prime}\boldsymbol{\omega}^*} + \frac{\boldsymbol{\omega}^{*\prime}\boldsymbol{R}^*(\boldsymbol{\eta})\boldsymbol{\omega}^*}{\boldsymbol{\omega}^{*\prime}\boldsymbol{\omega}^*}\right)$$

lies in the two expressions

$$\xi^2 := \frac{\boldsymbol{\omega}'\boldsymbol{R}(\boldsymbol{\eta})\boldsymbol{\omega}}{\boldsymbol{\omega}'\boldsymbol{\omega}} \quad \text{and} \quad \xi^{*2} := \frac{\boldsymbol{\omega}^{*\prime}\boldsymbol{R}^*(\boldsymbol{\eta})\boldsymbol{\omega}^*}{\boldsymbol{\omega}^{*\prime}\boldsymbol{\omega}^*}.$$

At first sight, the difference between ξ^2 and ξ^{*2}, and hence between MSFE and MSFE*, may seem trivial. This, however, is not so. First, while MSFE depends on the model selection procedure (for example, forward (specific-to-general) or backward (general-to-specific)), MSFE* is independent of the selection procedure. Secondly, while the eigenvalues of $\boldsymbol{R}(\boldsymbol{\eta})$ are not necessarily bounded, the eigenvalues of $\boldsymbol{R}^*(\boldsymbol{\eta})$ are always bounded, so that ξ^{*2} is always finite even when ξ^2 is infinite.[8] Thirdly, simple analytical expressions exist for the MSFE*, but not for MSFE. And finally, the 'optimal' WALS forecast can be applied quite easily in the case of MSFE*, but not in the case of MSFE.

We now compare the three procedures: forward, orthogonal, and Laplace. The forward pretest procedure was already discussed and applied; it is the standard procedure used in applied work. The orthogonal procedure first transforms the auxiliary regressors $\boldsymbol{Z}$ so that they become 'orthogonal' (in the sense that $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$), and then applies the standard pretest procedure to the transformed model. In this case it does not matter whether the

[8] In the forward pretest procedure, ξ^2 can become as large as we please by making $\boldsymbol{M}\boldsymbol{z}_i$ and $\boldsymbol{M}\boldsymbol{z}_j$ more and more correlated; see Danilov and Magnus (2001, Section 7).

pretest procedure is forward, backward, or something else; they all lead to the same result. Finally, the Laplace procedure is based on the transformed model, but it will use *all* auxiliary (transformed) regressors. The weights λ_i will determine how much weight is attached to each auxiliary regressor, essentially depending on the relevant t-statistic. The Laplace procedure can thus be viewed as a continuous version of the discrete (and hence inadmissible) pretest procedure.

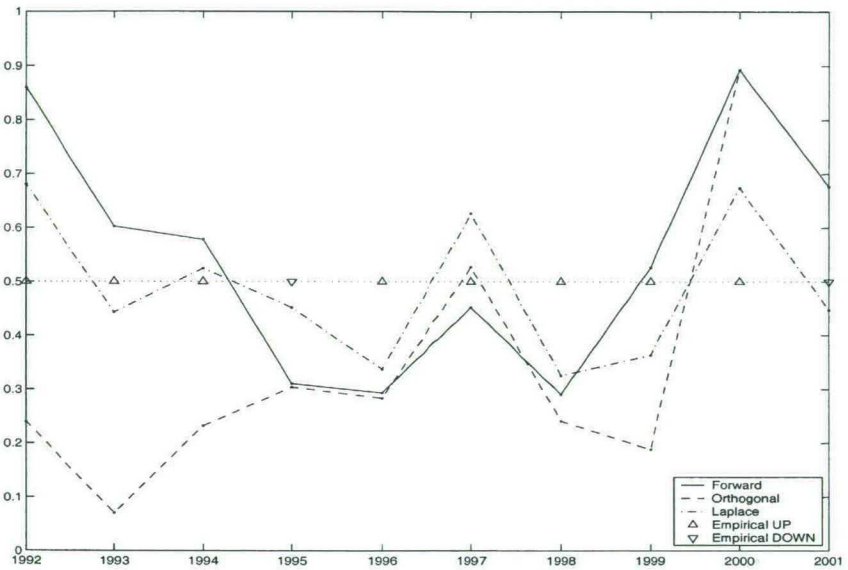

Figure 3. Forecast probabilities $\Pr(y_{n+1} > 0)$ for three procedures.

The main conclusion from Figure 3 is that none of the three procedures considered predict particularly well. The 2001 crash, for example, was only predicted by the Laplace procedure. The triangles depict the direction of the market: down in 1995 and 2001, up in the other eight years. Of the thirty predictions (10 years, 3 procedures), exactly one half were correct. For example, in 1992, pretest and Laplace predicted correctly, but orthogonal predicted incorrectly. In 1996 and 1998 the market went up, but all three procedures predicted that it would go down. It turns out that predicting the annual excess returns on common stocks is a very hazardous business.

This does not, however, imply that all three *procedures* are equally bad. Let us consider the MSFE of the WALS estimator, given in Theorem 1:

$$\begin{aligned}\text{MSFE} &= \sigma^2 \left(\boldsymbol{x}'_{n+1} (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_{n+1} + \boldsymbol{\omega}' \text{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}}) \boldsymbol{\omega} + 1 \right) \\ &= \sigma^2 \left(\psi_2(\boldsymbol{\eta}) + \psi_1^2(\boldsymbol{\eta}) + 1 \right).\end{aligned}$$

This expression depends on $\boldsymbol{\eta}$ (and σ^2), which is unknown. Following the same approach as before, we obtain a 95% bound for the MSFE as

$$\text{MSFE} < \sigma^2 \left(C_5(\hat{\boldsymbol{\eta}}) + 1 \right),$$

where

$$C_5(\hat{\boldsymbol{\eta}}) := \max_{\boldsymbol{\eta} \in \mathcal{H}(\hat{\boldsymbol{\eta}})} \left(\psi_2(\boldsymbol{\eta}) + \psi_1^2(\boldsymbol{\eta}) \right).$$

In Figure 4 we compare the bounds of the MSFE for the forward pretest, orthogonal, and Laplace procedures.

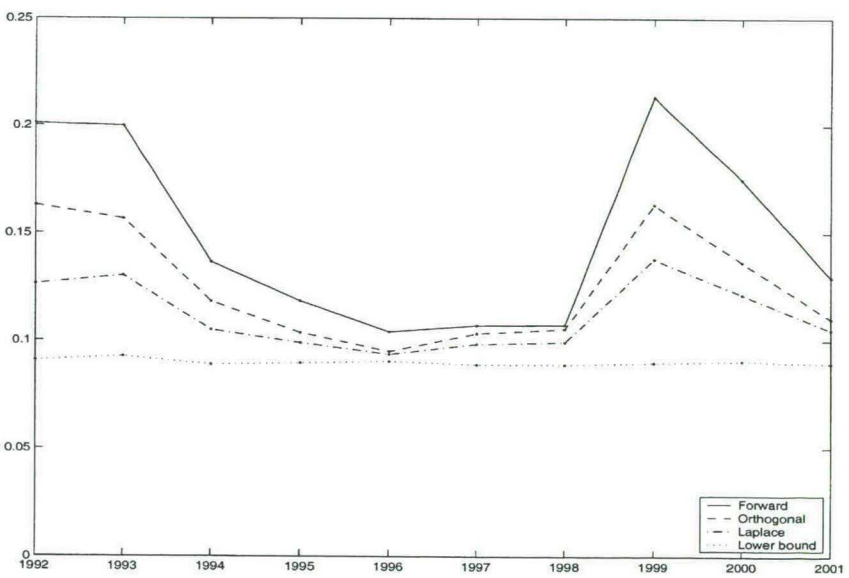

Figure 4. Upper bounds of MSFE, σ^2 estimated by (3.10).

Figure 4 shows convincingly the superiority of the Laplace estimator. Its MSFE bound is very much lower than for the pretest estimator, and uni-

formly so. Moreover, if we compare the bounds with their theoretical minimum

$$\sigma^2 \left(\boldsymbol{x}'_{n+1} (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_{n+1} + 1 \right),$$

then the difference between the procedures becomes even more pronounced. We also observe that the MSFE bounds vary significantly over time.

We thus conclude that — if our focus is forecasting rather than model selection — substantially better forecasts can be generated using the Laplace weights.

3.6 The effect of estimating σ^2

So far we derived the prediction intervals on the assumption that σ^2 is known, and only at the final stage did we substitute σ^2 by its estimate (3.10), based on the unrestricted model.

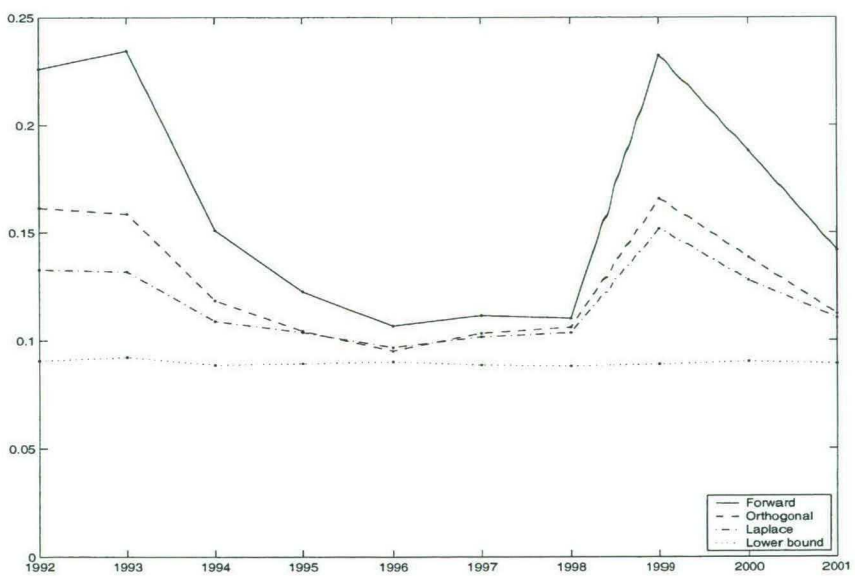

Figure 5. Upper bounds of MSFE, σ^2 estimated 'properly'.

We now want to treat σ^2 'properly', that is, we estimate it by the LS estimate of the selected sub-model

$$s^2_{(i)} = \frac{1}{n-k-m+r_i}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_{(i)} - \boldsymbol{Z}\boldsymbol{c}_{(i)})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_{(i)} - \boldsymbol{Z}\boldsymbol{c}_{(i)}),$$

and we take its distribution into account when selecting the model. There is no theoretical problem in doing the calculations, because the estimator for σ^2 will depend on $\boldsymbol{M}\boldsymbol{y}$, so that Theorem 1 still applies, but they are much more complicated and time-consuming. In Figure 5 we recalculate the MSFE-bounds of Figure 4, but now taking the estimation of σ^2 into proper account. As the plots show, the difference between Figures 4 and 5 is very small. This confirms the conclusion in Danilov (2002) that all qualitative (and most quantitative) results are not affected when we ignore the obvious fact that σ^2 is not known.

3.7 Concluding remarks

On the basis of our theoretical and empirical results, we conclude that taking explicit account of pretesting in assessing the properties of one-period-ahead forecasts is essential in econometrics, if we wish to be (or become) credible to policy makers and others.

We all know that we use the same data for model selection and forecasting (and estimation), that therefore pretesting takes place, and hence that the properties of forecasts (and estimators) are affected. This paper shows that it is possible to take pretesting into proper account, and that it matters. The conclusions of PT94 are much less robust than naive econometrics might imply, when the effects of pretesting are properly accounted for.

In addition, we show that an alternative exists to the (discontinuous, hence inadmissible) traditional pretest procedure, based on Laplace weights. These weights have optimal theoretical properties, and they appear to behave well in practice too.

Data appendix

We attempted to use the same data as in PT94 (Pesaran and Timmermann, 1994), but could not quite do so for four reasons. First, the data set used

by PT94 is not available now. We had access to the data used by PT95 (Pesaran and Timmermann, 1995); not, however, to the original data, but the data recently updated by the authors. Secondly, our data set extends to the year 2001, so that we had to employ a slightly different definition of the term premium TERM, since the 6-month commercial paper rate is no longer published by the Federal Reserve. Thirdly, we had no access to the CRSP (Center for Research in Security Prices) tapes, in particular not to the Fama-Bliss risk free rates files, that were used by Pesaran and Timmermann. Therefore an alternative source had to be used. Finally, various business cycle indicators employed in PT94 are in fact composite indices, subject to revisions and renormalizations. The indices that agree with the Citybase definition (used in PT94) end in November 1995, and a slightly different definition was employed afterwards. In this appendix we describe briefly how the data are constructed. Tables 1 and 2 provide the full data set employed.

dependent variable

The dependent variable ρ_t denotes the excess return in year t, and is defined by

$$\rho_t = \mathrm{NRSP}_t - \mathrm{I12}_{t-1},$$

where

$$\mathrm{NRSP}_t = \frac{\mathrm{PSP}_t - \mathrm{PSP}_{t-1} + \mathrm{DIVSP}_{t-1}}{\mathrm{PSP}_{t-1}}$$

denotes the annual rate of return on the SP 500 index, and $\mathrm{I12}_{t-1}$ denotes the 12-month T-bill rate on the last trading day of January in the year $t-1$.

The variable I12 is obtained from PT95, up to the year 1992. Later years are obtained from the H15 Federal Reserve Statistical Release, section Weekly Releases, Selected Interest Rates, Historical data, Treasury bills, Secondary market, 1-year, Business.[9]

The variable PSP denotes the nominal price index for the SP 500 portfolio at the close of the last trading day of January. Sources: PT95 (for the years 1955–1992) and DataStream (from 31 December 1964 up to 2001). We used the PT95 data set updated from DataStream where necessary.

DIVSP denotes the average nominal dividends per share for the SP 500 portfolio paid during the calendar year. It is constructed as DIVSP =

[9]See *http://www.federalreserve.gov/releases/h15/data/b/tbsm1y.txt.*

PSP × YSP, where YSP is defined below.

focus regressors

The first focus regressor is the constant term. In addition, we have three other focus regressors. The second is PI, the annual inflation rate, computed as $PI_t = \log(PPIAV_t/PPIAV_{t-1})$, where PPIAV denotes the annual average of the producer price index (PPI, finished goods). Source: website of the U.S. Department of Labor, Bureau of Labor Statistics, Series: Producer Price Index by Finished Goods (April 1947 to present).[10]

The third is DI3, the change in the 3-month T-bill rate, defined as the difference between the 3-month T-bill rate in January (I3:JAN) and the 3-month T-bill rate in October (I3:OCT) of the previous year, both measured at the last trading day of the month. Source: H15 Federal Reserve Statistical Release, section Weekly Releases, Selected Interest Rates, Historical data, Treasury bills, Secondary market, 3-month, Business.[11]

The fourth focus regressor is TERM, the term premium, defined as the difference between the 3-month financial paper rate (IF3:JAN) and I3:JAN. PT94 measure the term premium as the difference between the 6-month commercial paper rate (risky) and the 3-month T-bill rate (riskless) in January. Since the 6-month commercial paper rate does not exist after 1997, we use the 3-month financial paper rate instead. The 3-month financial paper rate data consist of two files, before September 1997 and after. Sources: H15 Federal Reserve Statistical Release, section Weekly Releases,Selected Interest Rates, Historical data, Finance paper placed directly (historical), 3-month, Monthly (1955–1997), and H15 Federal Reserve Statistical Release, section Weekly Releases, Selected Interest Rates, Historical data, Commercial paper (Financial), 3-month, Monthly (1997–2002).[12]

auxiliary regressors

We consider four auxiliary regressors. First, YSP, the dividend yield on the SP 500 portfolio, is defined as $YSP_t = DIVSP_{t-1}/PSP_t$. Sources: PT95,

[10] Available online at *www.bls.gov*.

[11] See *http://www.federalreserve.gov/Releases/h15/data/b/tbsm3m.txt*.

[12] See *http://www.federalreserve.gov/Releases/h15/data/m/hfp3m.txt* for the historical data and .../*fp3m.txt* for the recent data.

datafile (1955–1992), and DataStream (from January 1965 to present).

Secondly, DIP, the annual change in industrial production, is computed as $\text{DIP}_t = \log(\text{IPAV}_t/\text{IPAV}_{t-1})$, where IPAV is the 12-month average of the industrial production index (IP). Source: on-line database of the Federal Reserve Bank of St.-Louis.[13] The data are monthly, seasonally adjusted, and range from January 1940 to August 2001. The data series is an index, base year 1992.

Thirdly, PER, the price-earnings ratio for the SP 500 index, is the ratio of the price of stock to the earnings of companies per unit of stock. We have two sources for these variables, one from PT95, the other from DataStream. (Note that PT95 give the earnings-price ratio, rather than the price-earnings ratio.) DataStream use the annualized price-earnings ratio.

Finally, DLEAD denotes the annual change in the leading business cycle indicator, and is defined as $\text{DLEAD}_t = \log(\text{LEAD}_t/\text{LEAD}_{t-1})$. Here, LEAD is the 12-month average of a composite of 11 leading business cycle indicators. The leading indicator LEAD is taken from the data set BCIH-01.dat (composite indexes), distributed by BCI Data Manager (January 1948 to November 1995).[14] For more recent data we extend the series as follows. We take the 'updated series' from the Economagic website.[15] This series is, however, calculated using a slightly different definition and base year. Therefore, we regress the old series on the updated series over the period where they overlap ($R^2 = 0.99$), and use the intercept and slope estimates and the values of the updated series to predict the missing years of the old series.

[13] See *www.stls.frb.org*, which in turn refers to the Federal Reserve Board, Washington, D.C.

[14] See *http://www.wfu.edu/~cottrell/bci/Software.html*.

[15] See *http://www.economagic.com/em-cgi/data.exe/feddal/jlead*.

year	ρ_t	PI_{t-2}	$DI3_{t-1}$	$TERM_{t-1}$
1956	0.2281	0.0022	0.12	0.13
1957	0.0365	0.0022	0.17	0.43
1958	-0.0572	0.0275	0.26	0.26
1959	0.3504	0.0366	-2.05	1.68
1960	0.0054	0.0221	0.09	0.36
1961	0.0997	-0.0020	-0.01	0.91
1962	0.1191	0.0083	0.16	0.44
1963	-0.0403	-0.0002	0.45	0.32
1964	0.1679	0.0032	0.21	0.25
1965	0.1311	-0.0035	0.02	0.33
1966	0.0511	0.0037	0.32	0.14
1967	-0.0858	0.0175	0.57	0.20
1968	0.0526	0.0308	-0.71	0.99
1969	0.0953	0.0122	0.32	0.59
1970	-0.2049	0.0286	0.71	-0.05
1971	0.0847	0.0359	0.87	0.28
1972	0.0730	0.0341	-1.69	0.93
1973	0.1054	0.0297	-0.96	0.61
1974	-0.2000	0.0306	0.92	-0.12
1975	-0.2344	0.0876	0.12	0.50
1976	0.2972	0.1425	-2.12	1.65
1977	-0.0034	0.1024	-0.78	0.43
1978	-0.1354	0.0430	-0.17	-0.08
1979	0.1058	0.0630	0.24	0.27
1980	0.1011	0.0753	0.54	0.81
1981	0.0726	0.1054	-0.12	-0.04
1982	-0.1544	0.1263	1.88	-0.10
1983	0.1283	0.0885	-0.23	0.04
1984	0.0863	0.0393	0.20	-0.14
1985	0.0500	0.0160	0.38	0.19
1986	0.1346	0.0207	-0.96	-0.24
1987	0.2575	0.0086	-0.22	0.55
1988	-0.0890	-0.0134	0.40	-0.01
1989	0.1278	0.0203	0.37	0.98
1990	0.0546	0.0247	1.03	0.39
1991	0.0011	0.0500	-0.03	0.16

Table 1a. Dependent variable and focus regressors, 1956–1991.

year	YSP_{t-1}	DIP_{t-1}	PER_{t-1}	$DLEAD_{t-2}$
1956	0.0510	136.2792	-0.0583	0.9986
1957	0.0398	148.1861	0.1204	1.0655
1958	0.0382	157.7279	0.0423	0.9949
1959	0.0391	155.4372	0.0141	0.9678
1960	0.0387	190.3797	-0.0670	1.0000
1961	0.0311	208.2837	0.1129	1.0492
1962	0.0348	202.3910	0.0223	0.9823
1963	0.0300	258.8171	0.0066	1.0232
1964	0.0333	212.3313	0.0800	1.0189
1965	0.0318	216.9554	0.0595	1.0235
1966	0.0297	223.2941	0.0652	1.0266
1967	0.0298	215.8158	0.0947	1.0247
1968	0.0333	185.6376	0.0848	1.0034
1969	0.0320	206.3934	0.0214	0.9931
1970	0.0307	210.8557	0.0541	1.0219
1971	0.0318	186.6205	0.0454	1.0045
1972	0.0382	198.8993	-0.0335	0.9573
1973	0.0318	222.2841	0.0136	1.0329
1974	0.0285	222.3302	0.0925	1.0477
1975	0.0301	161.8866	0.0781	1.0108
1976	0.0428	106.8587	-0.0151	0.9431
1977	0.0435	142.4935	-0.0916	0.9659
1978	0.0376	130.9383	0.0876	1.0695
1979	0.0446	105.7509	0.0783	1.0165
1980	0.0516	101.5938	0.0570	1.0108
1981	0.0526	91.4693	0.0327	0.9871
1982	0.0510	101.5335	-0.0280	0.9620
1983	0.0499	96.4215	0.0162	1.0023
1984	0.0575	115.2959	-0.0552	0.9809
1985	0.0440	148.3745	0.0366	1.0746
1986	0.0457	124.5979	0.0857	1.0182
1987	0.0419	155.8990	0.0163	1.0000
1988	0.0345	209.0400	0.0112	1.0252
1989	0.0303	204.4862	0.0453	1.0235
1990	0.0349	150.9822	0.0444	1.0010
1991	0.0326	164.9125	0.0178	0.9960

Table 2a. Auxiliary regressors, 1956–1991.

year	ρ_t	PI_{t-2}	$DI3_{t-1}$	$TERM_{t-1}$
1992	0.1580	0.0481	-0.92	0.73
1993	0.0616	0.0214	-0.98	0.15
1994	0.0852	0.0122	-0.06	0.42
1995	-0.0269	0.0124	-0.07	0.15
1996	0.3045	0.0063	0.80	0.27
1997	0.2181	0.0190	-0.41	0.34
1998	0.2147	0.0258	-0.01	0.30
1999	0.2703	0.0041	-0.01	0.38
2000	0.0323	-0.0088	0.14	0.44
2001	-0.0360	0.0180	0.56	0.28

Table 1b. Dependent variable and focus regressors, 1992–2001.

year	YSP_{t-1}	DIP_{t-1}	PER_{t-1}	$DLEAD_{t-2}$
1992	0.0349	186.9279	-0.0021	0.9880
1993	0.0326	264.9310	-0.0202	0.9868
1994	0.0297	291.4800	0.0309	1.0103
1995	0.0277	278.0400	0.0343	1.0071
1996	0.0283	205.2000	0.0527	1.0293
1997	0.0255	211.4400	0.0469	0.9910
1998	0.0218	257.5200	0.0447	1.0117
1999	0.0176	288.6000	0.0650	1.0195
2000	0.0147	391.6800	0.0466	1.0169
2001	0.0125	304.9200	0.0406	1.0215

Table 2b. Auxiliary regressors, 1992–2001.

Chapter 4

Estimation of the mean of a univariate normal distribution when the variance is not known

4.1 Introduction

Let x be a single observation from a univariate normal distribution with unknown mean θ and unknown variance σ^2, that is $x \sim \mathrm{N}(\theta, \sigma^2)$. Suppose also, that an estimator of σ^2 is available, namely s^2, that is independent of x and such that $s^2\nu/\sigma^2$ has a χ^2_ν distribution, where ν is known. In this article we consider the problem of estimating θ in some optimal manner.

The stated problem arises from the following practical situation. Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\gamma\boldsymbol{z}+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. The difference between $\boldsymbol{X}$ and $\boldsymbol{z}$ is that $\boldsymbol{X}$ contains regressors that always have to be in the model, while $\boldsymbol{z}$ (also called the auxiliary regressor) may or may not be in the model. We want to estimate parameter $\boldsymbol{\beta}$ in the 'best' possible way.

The problem of optimal estimation of $\boldsymbol{\beta}$ has a long history. Early work on pretesting goes back to Berkson (1942) and Bankroft (1944). Huntsberger (1955) explicitly writes out the pretest estimator as a continuous weighted average of the restricted and unrestricted estimators, and proposes an alternative estimator. Feldstein (1973) is concerned with regression estimation when regressors are highly correlated. Admissibility issues are discussed in Blyth (1951), Farrell (1968), Brown (1971), and Berger (1976). A review

of the early literature is provided in Judge and Bock (1978). Sawa and Hiromatsu (1973) consider the pretest estimator using the minimax regret criterion, while Toyoda and Wallace (1976) use the average minimum risk criterion. The minimax regret approach is also used by Droge and Georg (1995) in obtaining adaptive least-squares regression estimates. Roehrig (1984) finds an expression for the mean squared error of the pretest estimator. In a similar way Magnus and Durbin (1999) derive moments of the general weighted-average least-squares estimator. Magnus (2002) introduces the neutral Laplace WALS estimator in a regression context. Recent developments in theory and practice of pretesting can be found in Giles and Giles (1993), Chatfield (1995), and Magnus (1999).

The rest of paper is organized as following. In section 2 we introduce notation and explain the basic setup. In section 3 we reconsider properties of WALS estimators, such as admissibility, risk and regret in the situation when σ is unknown. In section 4 we consider the performance of the neutral Laplace estimator relative to the usual pretest estimator. Section 4 concludes. An appendix contains proofs of all results.

4.2 Notation and setup

We consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n), \tag{4.1}$$

where $\boldsymbol{y}(n\times 1)$ is the vector of observations, $\boldsymbol{X}(n\times k)$ and $\boldsymbol{z}(n\times 1)$ are matrices of nonrandom regressors, $\boldsymbol{\varepsilon}(n \times 1)$ is a vector of disturbances, and $\boldsymbol{\beta}$ and γ are unknown nonrandom parameters. Suppose that the matrix $(\boldsymbol{X} : \boldsymbol{z})$ has full column-rank. Let $\boldsymbol{b}_u$ and $\hat{\gamma}$ denote the ordinary least-squares estimators of $\boldsymbol{\beta}$ and γ in model (4.1), and $\boldsymbol{b}_r$ the ordinary least-squares estimator in model (4.1) under the restriction $\gamma = 0$. Denote $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\theta = \gamma/\sqrt{\boldsymbol{z}'\boldsymbol{M}\boldsymbol{z}}$ and $\hat{\theta} = \hat{\gamma}/\sqrt{\boldsymbol{z}'\boldsymbol{M}\boldsymbol{z}}$. The weighted-average least-squares estimator, introduced in Magnus and Durbin (1999), is defined as

$$\boldsymbol{b} = \lambda\boldsymbol{b}_u + (1-\lambda)\boldsymbol{b}_r, \quad \lambda = \lambda(\hat{\theta}, s^2),$$

where $s^2 = y'My/(n-k-1)$ is the least-squares estimator of σ^2. The equivalence theorem in Magnus and Durbin (1999) states that the mean

squared error of the WALS estimator of $\boldsymbol{\beta}$ can be represented as

$$\mathsf{MSE}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \mathsf{MSE}(\lambda(\hat{\theta}, s^2)\hat{\theta})\, \boldsymbol{q}\,\boldsymbol{q}',$$

where $\boldsymbol{q} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{M}\boldsymbol{z})^{-1/2}$. Therefore we can say that the problem of estimation of $\boldsymbol{\beta}$ in regression model (4.1) is equivalent to the problem of estimation of the parameter θ by one bivariate observation $(\hat{\theta}, s^2)$. We call the first problem *the regression problem* and second problem *the auxiliary problem*. For determining the optimal $\boldsymbol{b}$ (in the mean squared error sense) we need to find a function $\lambda(\hat{\theta}, s^2)$ which provides the optimal estimator of θ. Consider the mean squared error of $\lambda(\hat{\theta}, s^2)\hat{\theta}$, that is,

$$\mathsf{MSE}(\lambda(\hat{\theta}, s^2)\hat{\theta}) = \mathsf{E}(\lambda(\hat{\theta}, s^2)\hat{\theta} - \theta)^2,$$

where the expectation on the right hand side is taken with respect to the product of two independent distributions: $\mathrm{N}(\theta, \sigma^2)$ and χ_ν. In Magnus (2002) the problem of finding the optimal λ was considered under the assumption that σ is known. He used a λ-function of particular form: $\lambda(\hat{\theta}, s^2) = \Lambda(\hat{\eta})$, where $\hat{\eta} = \hat{\theta}/\sigma$. Then,

$$\mathsf{MSE}(\lambda(\hat{\theta}, s^2)\hat{\theta}) = \sigma^2\mathsf{E}(\Lambda(\frac{\hat{\theta}}{\sigma})\frac{\hat{\theta}}{\sigma} - \frac{\theta}{\sigma})^2 = \sigma^2\mathsf{E}(\Lambda(\hat{\eta})\hat{\eta} - \eta)^2,$$

where $\eta = \theta/\sigma$. Therefore the regression problem with known σ is equivalent to the estimation of the parameter η by one observation from the $\mathrm{N}(\eta, 1)$ distribution.[1] In general, however, σ is not known. We estimate σ by the usual least-squares estimator, $s = \sqrt{y'My/(n-k-1)}$. The expression $\hat{\theta}/s$ is then interpreted as the usual regression t-statistic. The WALS estimator based on

$$\lambda(\hat{\theta}, s^2) = \Lambda(\hat{\theta}/s) \tag{4.2}$$

generalises in a natural way the usual pretest estimator. The mean squared error of $\lambda(\hat{\theta}, s^2)\hat{\theta}$ is now given by

$$\mathsf{MSE}(\lambda(\hat{\theta}, s^2)\hat{\theta}) = \sigma^2\mathsf{E}(\Lambda(\frac{\hat{\theta}}{s})\frac{\hat{\theta}}{\sigma} - \frac{\theta}{\sigma})^2 = \sigma^2\mathsf{E}(\Lambda(\frac{\hat{\eta}}{v}\sqrt{\nu})\hat{\eta} - \eta)^2, \tag{4.3}$$

where $v = \frac{s}{\sigma}\sqrt{\nu}$, $\nu = n - k$. In expression (4.3), the statistics v and $\hat{\eta}$ are independent and distributed as χ_ν and $\mathrm{N}(\eta, 1)$ respectively. We see that the

[1] In Magnus (2002) this is called *the* $\mathrm{N}(\eta, 1)$ *problem*.

risk function depends on the unknown parameter σ, which is only a scale parameter. Therefore we can write the mean squared error as

$$\mathsf{MSE}(\lambda(\hat{\theta}, s^2)\hat{\theta}) = \sigma^2 R(\eta, \Lambda),$$

where the standardized risk function R is defined as

$$R(\eta, \Lambda) = \mathsf{E}_\eta \mathsf{E}_\nu (\Lambda(x\sqrt{\nu}/v)x - \eta)^2, \tag{4.4}$$

with $x \sim \mathrm{N}(\eta, 1)$, $v \sim \chi_\nu$, independent of x, and E_η denotes expectation with respect to $\mathrm{N}(\eta, 1)$, and E_ν with respect to the χ_ν distribution. The function R can be interpreted as the mean squared error of the WALS estimator in the case where $\sigma = 1$. Now denote

$$R_v(\eta, \Lambda) = \mathsf{E}_\eta (\Lambda(x\sqrt{\nu}/v)x - \eta)^2, \tag{4.5}$$

so that then $R(\eta, \Lambda) = \mathsf{E}_\nu R_v(\eta, \Lambda)$. The function $R_v(\eta, \Lambda)$ can be interpreted as the risk function for the problem with known variance but with different Λ, specifically $\Lambda_v(x) = \Lambda(x\sqrt{\nu}/v)$.

4.3 WALS estimation in auxiliary problem with unknown variance

We wish to investigate properties of the estimator of η, based on one bivariate observation (x, s), where $x \sim \mathrm{N}(\eta, 1)$, $s = w\sigma/\sqrt{\nu}$, $w \sim \chi_\nu$, independent on x The estimator takes the form

$$t(\Lambda, x, s) = \Lambda(x/s)x. \tag{4.6}$$

Note, that the weight $\Lambda(x/s)$ depends not only on x but also on the independent statistic s. This additional randomness is caused by the necessity to estimate the nuisance parameter σ. We will assume that Λ satisfies the regularity conditions R1, i.e.

a. $0 \leq \Lambda(x) \leq 1$, for all x,

b. $\Lambda(-x) = \Lambda(x)$, for all x,

c. Λ is nondecreasing on $[0, \infty]$,

d. Λ is continuous except possibly on a set of measure zero.

These conditions allow us to interpret $\Lambda(x/s)x$ as a shrinkage estimator. There is an apparent parallel in investigating the properties of (4.6) for known and unknown σ. In particular, all notions of admissibility of the estimators can be reformulated straightforwardly by considering the new risk function,

$$R(\eta, \Lambda) = \mathsf{E}(t(\Lambda, x, s) - \eta)^2,$$

but the actual investigation of admissibility and regret requires some care. Our main question is whether the conclusions made for the WALS estimator with known variance are still true (or almost true) in the case with unknown variance. In Magnus (2002) several important one-parametric classes of WALS estimators were considered. The *normal Bayes* estimator is defined as $t^{(1)} = x/(1+c)$. The usual pretest estimator is defined by choosing Λ as

$$\Lambda_c^{(2)}(x) = \begin{cases} 0 & \text{if } |x| \leq c, \\ 1 & \text{if } |x| > c, \end{cases} \tag{4.7}$$

and the class of all pretest estimators that have $0 < c < \infty$ is denoted as $\mathcal{L}^{(2)}$. The Laplace estimator corresponds to the Λ-function of the form

$$\Lambda_c^{(3)}(x) = 1 - \frac{h(x)}{x}c, \tag{4.8}$$

where

$$h(x) = \frac{1 - e^{2cx}\psi(x)}{1 + e^{2cx}\psi(x)}, \quad \psi(x) = \frac{\Phi(-x-c)}{\Phi(x+c)},$$

and the class of all Laplace estimators with positive finite c is denoted as $\mathcal{L}^{(3)}$. The 'neutral' Laplace estimator corresponds to $c = 0.6931$. Finally the Burr estimator was defined as

$$\Lambda_c^{(4)}(x) = \begin{cases} 0 & \text{if } |x| \leq c, \\ 1 - \frac{c}{|x|} & \text{if } |x| > c. \end{cases} \tag{4.9}$$

The corresponding class of Burr estimators is denoted by $\mathcal{L}^{(4)}$.

The natural starting point is admissibility. The usual estimator is defined by $\Lambda = 1$. Since this Λ does not depend on s at all, we can expect that little is changed in the properties of the resulting estimator. Indeed, Theorem 3

shows that the usual estimator is unbiased, admissible and has constant risk equal to 1. Similar considerations (Theorem 4) show that the normal Bayes estimators are admissible for any $0 \leq c < \infty$. In the auxiliary problem with known σ, the pretest estimator is proved to be $\mathcal{L}^{(2)}$-admissible. Theorem 7 shows that the pretest estimator for each value of c is $\mathcal{L}^{(2)}$-admissible also when σ is unknown. A similar result holds for the Burr estimator (Theorem 8). This simply reflects the fact that in each class no estimator dominates an other. Moreover their risks are bounded and converge pointwise to 1 as c tends to infinity. For the Laplace estimator it is possible to prove a stronger property, namely that the Laplace estimator is admissible (not just $\mathcal{L}^{(3)}$-admissible) even when σ is unknown (Theorem 9). We can also establish $\mathcal{L}^{(1)}$-admissibility of the pretest estimator and $\mathcal{L}^{(2)}$-admissibility of the Burr estimator.

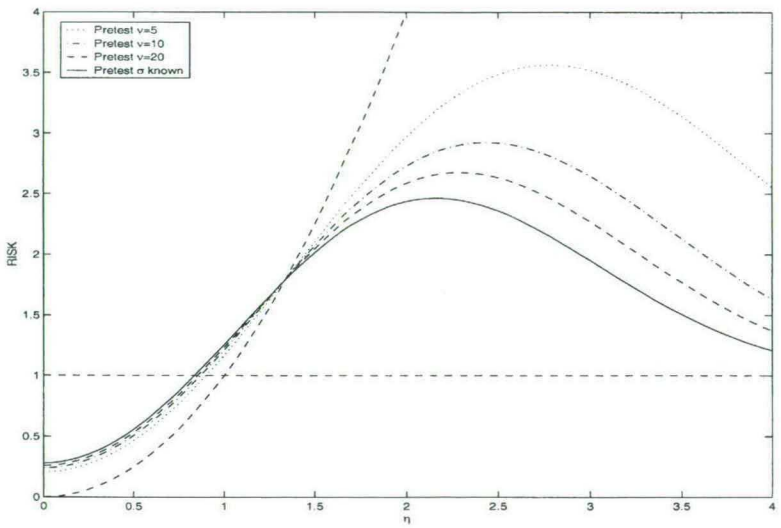

Figure 1. Risk function of the ordinary 5% pretest estimator for various values of ν.

It is well known (see e.g. Judge and Bock (1978)) that the risk function of the pretest estimator depends on ν. Figure 1 shows that this is indeed the case. In the figure we plot the risk functions of the ordinary 5% pretest

estimator[2], together with the risk functions of the restricted ($\Lambda = 0$) and the unrestricted ($\Lambda = 1$) estimators. First of all note that as ν grows, the risk profiles converge to some limiting function, say $R_\infty(\theta)$. This function is nothing more than the risk function of the pretest estimator in the auxiliary problem with known σ. Indeed, the conditional risk function (4.5) satisfies

$$R_v(\theta, \Lambda) = R_v(\theta, \Lambda_c) = R_1(\theta, \Lambda_{\frac{cv}{\sqrt{\nu}}}).$$

Since $v/\sqrt{\nu} \to 1$ a.s., it follows that $R(\eta, \Lambda) = \mathsf{E}_\nu R_v(\eta, \Lambda) \to R_1(\theta, \Lambda_c)$, uniformly in θ. It is also quite obvious (and Figure 1 confirms this) that the risk functions of the restricted and the unrestricted estimators do not depend on ν. We know that $c = c_\nu = T_\nu^{-1}(0.975)$ grows as ν decreases. Therefore for small values of ν the 5% pretest estimator behaves more and more like the restricted estimator. This explains the fact that as ν decreases, risk profiles become smaller for small η and larger for large η. It appears, however, that risks depend very little on ν around the point $\eta = 1.3370$, where the graphs $R_\infty(\theta)$ and θ^2 intersect.[3] Similarly, there is very little dependence on ν around the point $\eta = 1$. Hence in this important region the pretest estimator still remains worse than both the restricted and the unrestricted estimator. The maximum of the risk for the ordinary pretest estimator becomes larger when ν decreases. For $\nu = 5$ the maximum risk is 45% higher than for the case of known σ. Therefore for small samples the usual 5% pretest estimator seems to perform even worse than for the case of known σ. In contrast, the neutral Laplace estimator reveals considerable robustness. In Figure 2 we plot the risk functions of the neutral Laplace estimator for different values of ν.

Figure 2 shows that the risk of the Laplace estimator depends remarkably little on ν.[4] Comparing with the case of known σ, the risk for small ν is slightly higher when $|\eta| < 2$, but outside this region the situation is reversed. Of course, the limiting value for $\eta \to \infty$ is the same and equal to $1 + c^2$.

[2]Recall that for 5% pretest estimator parameter c is determined as $c = T_\nu^{-1}(0.975)$, where T_ν is c.d.f. of t-distribution with ν degrees of freedom.

[3]Actually, the dependence of R on ν does not disappear completely but only becomes very small.

[4]Note that in Figure 2 the scale on vertical axis is larger than in Figure 1. If we would not do this, the reader would not see *any* difference in risk profiles for different ν.

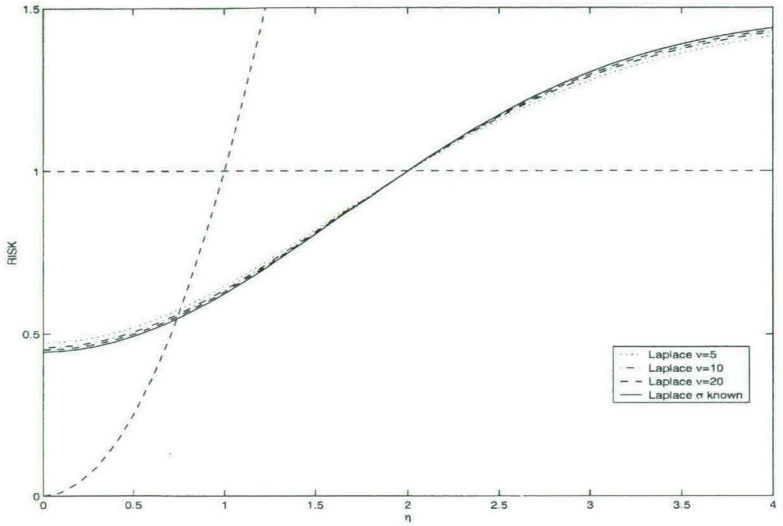

Figure 2. Risk of the neutral Laplace estimator as function of η for various values of ν.

One of the well-accepted approaches in finding an optimal Λ is the minimax regret approach (Savage (1951), Chernoff and Moses (1959), Sawa and Hiromatsu (1973)). An optimal Burr estimator was obtained in Magnus (2000) as the minimax regret solution within a specific subclass. Magnus also used regret to characterise optimality of the Laplace estimator. In order to apply this approach for our case we need to find the regret function. Regret is defined as the difference between risk of the estimator and minimal risk in a given class of estimators. Theorem 11 establishes that the minimal risk of the WALS estimator within all $R1$-regular $\mathcal{L}^{(0)}$ Λ-s is the same as in the problem with known σ, that is,

$$\inf_{\lambda \in \mathcal{L}^{(0)}} R(\eta, \lambda) = \frac{\eta^2}{1+\eta^2}.$$

Therefore, the regret function for a particular WALS estimator is defined as

$$r_0(\eta, \Lambda) = R(\eta, \Lambda) - \frac{\eta^2}{1+\eta^2}.$$

Now, let us see if and how the regret properties of the neutral Laplace estimator change in the problem with unknown σ.

ν	5	10	15	20	40	∞
min r_0	0.1183	0.1099	0.1069	0.1053	0.1030	0.1006
max r_0	0.5036	0.5075	0.5091	0.5099	0.5112	0.5127
min R	0.4725	0.4582	0.4536	0.4513	0.4479	0.4446
max R	1.4805	1.4805	1.4805	1.4805	1.4805	1.4805

Table 1. Extreme values for risk and regret of the neutral Laplace estimator.

In Table 1 we gathered the extreme values for risk and regret of the Laplace estimator for different values of ν. The last column represents the case of known σ. We see that minimum risk is about 6% higher for small ν, but that the maximum risk is the same. In contrast, maximum regret is only about 2% lower for small ν, but minimum regret is about 18% higher. (This is a direct consequence of the fact that the lower bound of risk of the WALS estimator is the same for all η.) Each row converges quickly to their limits when ν grows. This limiting value coincides of course with the value for the case of known σ.

Summarising we can say that the neutral Laplace estimator, developed for the case of known σ, performs exceptionally good. The difference in properties caused by estimation of σ is small. In terms of mean squared error this difference does not exceed 5% for the relevant range of ν. Therefore we recommend to use the neutral Laplace estimator in practice.

Nevertheless one potential resource for improvement still remains. It is possible to apply, for example, the optimal minimax regret Burr estimator in the new setup. Since the regret function is different for each ν, this optimal solution will take into account this dependence. *A priori* it is not clear, how much we can gain by applying this approach. Figure 3 classifies the situation.

In the Figure 3 we plotted risk profiles of the minimax regret Burr estimator for various values of ν. Comparing Figure 3 with Figure 2 we can only conclude that the potential gain is marginal. Hence, adapting WALS estimators via the minimax regret principle to each ν does not seems to be a productive idea. The Laplace solution still appears to be the best.

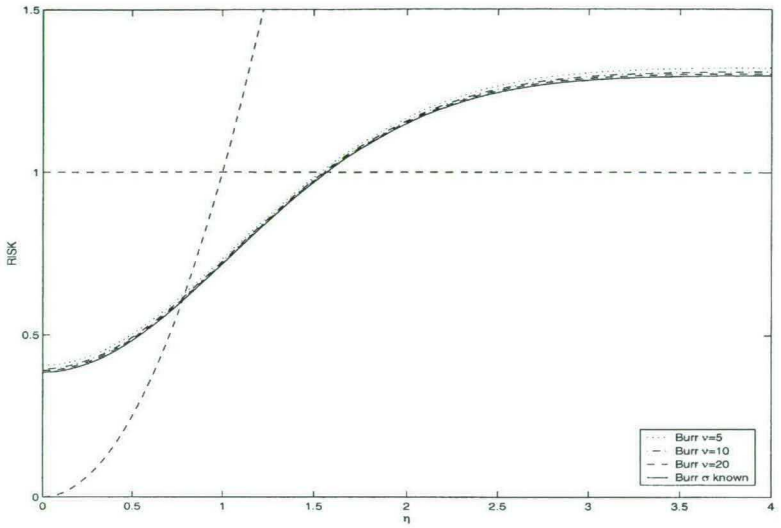

Figure 3. Risk of the optimal minimax regret Burr estimator as a function of η for various values of ν.

4.4 Relative efficiency of the Laplace estimator

In previous sections we investigated properties of several WALS estimators in the auxiliary problem. We have established several important properties concerning admissibility and risk of these estimators. Our general conclusion was that we can use the optimal solution obtained in Magnus (2002), because there is only little difference in risk and regret values obtained for known and unknown σ. However, our original problem concerned the estimation of a focus regressor in the partitioned regression problem. We now compare the performance of the Laplace and the pretest estimator in a real regression problem. For this purpose we consider the ratio of the mean squared errors of the ordinary pretest estimator and the neutral Laplace estimator. More precisely, let us consider the problem of estimating a general linear combination of the parameters $\boldsymbol{\beta}$, say $\boldsymbol{\omega}'\boldsymbol{\beta}$, where $\boldsymbol{\omega}$ is a known $k \times 1$ vector. The

mean squared error of $\boldsymbol{\omega}'\boldsymbol{b}$, according to the equivalence theorem, is

$$\begin{aligned}\mathsf{MSE}(\boldsymbol{\omega}'\boldsymbol{b}) &= \sigma^2(\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega} + \boldsymbol{\omega}'\boldsymbol{q}\boldsymbol{q}'\boldsymbol{\omega}R(\eta)) \\ &= \sigma^2\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega}(1 + q_0^2 R(\eta)),\end{aligned}$$

where

$$q_0^2 = \frac{(\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}z)^2}{\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega}\ z'\boldsymbol{M}\boldsymbol{z}},$$

and $R(\eta)$ is a generic notation for the risk function of the WALS estimator in the auxiliary problem. Therefore the ratio of mean squared errors of the neutral Laplace and pretest estimators is

$$G = \frac{1 + q_0^2 R^P(\eta)}{1 + q_0^2 R^L(\eta)},$$

where $R^L(\eta)$ denotes the risk of the neutral Laplace estimator, and $R^P(\eta)$ the risk of the ordinary pretest estimator. Values of G larger than 1 correspond to the region where the Laplace estimator performs better than the ordinary pretest, and vise versa.

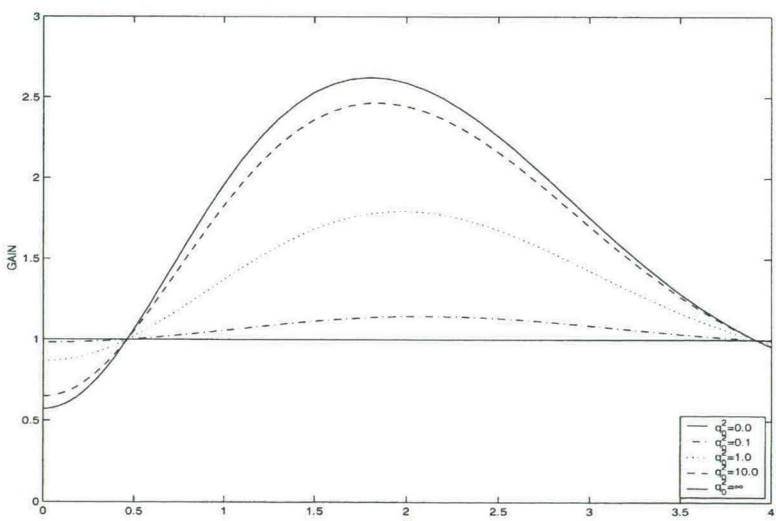

Figure 4. Relative efficiency of the neutral Laplace estimator compared with the usual 5% pretest estimator, $\nu = 20$.

Figures 4 illustrates the behavior of the gain profile G for different values of q_0^2 as a function of η. We set ν equal to 20. First, we see that for values of q_0^2 close or equal to 0 we have little or no gain compared with the usual pretest case. The Laplace estimator performs better in the interval [0.46, 3.93]. For very small and very large values of η, the Laplace estimator is slightly worse than the usual pretest estimator. Actually, this superiority interval will be different for different ν. Table 2 shows the nature of this dependence.

ν	1	5	10	15	20	30	40	∞
x_*	0.7732	0.5622	0.4985	0.4743	0.4615	0.4483	0.4416	0.4207
x^*	28.4806	5.0971	4.2447	4.0204	3.9174	3.8200	3.7732	3.6405

Table 2. Superiority interval of Laplace estimator as function of ν.

We see from the Table 2 that the superiority interval shrinks and moves to the right as ν grows. Both lower and upper boundaries of the interval tend monotonically to their limiting values 0.4207 and 3.6405 respectively.[5] Note again, that the Laplace estimator always performs better in the region of moderately large values of η, including the important neighborhood of point $\eta = 1$. Moreover, for small ν the superiority interval for the Laplace estimator is larger than for large ν.

4.5 Conclusion

In the current article we discussed issues connected with the practical application of the neutral Laplace weighted-average least-squares estimator, introduced in Magnus (2002), relaxing the assumption that the variance of innovation is known. We found that properties of the Laplace estimator are surprisingly similar whether or not σ is known. Important properties of the Laplace solution such as admissibility, bounded risk and small regret values still hold in the new setup. Moreover, a comparison of the Laplace and ordinary pretest estimators shows that there is a large interval where the Laplace estimator performs significantly better than the ordinary pretest estimator.

[5]These limiting values coincide of course with the lower and upper bounds for the superiority interval in the case when σ is known.

The superiority interval is larger for small sample size (more exactly, for small values of ν), but remains rather large even asymptotically. On the base of these results we recommend the neutral Laplace solution for use in practical situations.

Applicability of the Laplace estimator is not restricted to the estimation case. Simple calculations (not reproduced here) show that a relation, similar to the Magnus and Durbin (1999) Equivalence Theorem, holds also for the mean squared error of the out-of-sample regression prediction, and therefore the same optimal λ function can be used for prediction. Applicability of the Laplace estimator is also not restricted to one auxiliary regressor. Our preliminary investigations show that there are no difficulties in applying optimal WALS estimator in a regression problem with several auxiliary regressors, if these regressors are orthogonal in some sense. If not, then a non-degenerate linear transformation of the auxiliary regressors is necessary to achieve orthogonality.

4.6 Appendix

This appendix contains a set of results about risk, admissibility and regret of various WALS estimators in the auxiliary problem with unknown σ. Admissibility is investigated by application of standard methods of decision theory (see e.g. Berger (1985)). First, let us clarify our notation. Our decision rule δ is a function of two arguments $\delta = \delta(x, v) = \Lambda(x\sqrt{\nu}/v)x$. The loss function is quadratic $L(\delta, \eta) = (\delta - \eta)^2$, and the risk function is $R(\eta, \delta) = \mathsf{E}_\eta \mathsf{E}_v(\delta(x, v) - \eta)^2$, where expectation on the right hand side is taken with respect to the product of two independent distributions: $x \sim \mathrm{N}(\theta, \sigma^2)$ and $v \sim \chi_\nu$. We need the following auxiliary result that establishes continuity of the risk function.

Lemma 1 The risk function

$$R(\eta, \delta) = \mathsf{E}_\eta \mathsf{E}_\nu(\Lambda(x\sqrt{\nu}/v)x - \eta)^2$$

is continuous in the region $C = \{\eta : |\eta| < 1\}$.

Proof: By definition,

$$R(\eta,\delta) = \int_{\infty}^{\infty}\int_{0}^{\infty} (\Lambda(x\sqrt{\nu}/v)x - \eta)^2 \phi_{\eta,1}(x) f_\nu(v)\,\mathrm{d}x\,\mathrm{d}v$$
$$= \int_{\infty}^{\infty}\int_{0}^{\infty} (\Lambda(\frac{(y+\eta)\sqrt{\nu}}{v})(y+\eta) - \eta)^2 \phi_{0,1}(y) f_\nu(v)\,\mathrm{d}y\,\mathrm{d}v.$$

Therefore it is enough to prove that

$$\int_{\infty}^{\infty}\int_{0}^{\infty} \Lambda^n(\frac{(y+\eta)\sqrt{\nu}}{v}) y^m \phi_{0,1}(y) f_\nu(v)\,\mathrm{d}y\,\mathrm{d}v, \tag{4.10}$$

for $n = 0, 1, 2$ and $m = 0, \ldots, n$ are continuous in C. This is the classical case of continuity of Lebegue integral with respect to a parameter (see e.g. Kolmogorov and Fomin (1957) or Apostol (1974)). The only nontrivial thing to show is the existence of an integrable majorant function, that is $|y|^m \phi_{0,1}(y) f_\nu(v)$. This function is integrable because the normal distribution has moments of all orders, and it is a majorant for the integrand in (4.10) because $|\Lambda(u)| \leq 1$, by regularity condition R1. ‖

Theorem 2 (Brown-Farrell) The decision rule δ is admissible if there exists a sequence of finite nonnegative measures $\{G_n\}$ such that $G_n(C) \geq 1$ and

$$r(G_n, \delta) - r(G_n, \delta_{G_n}) \to 0, \text{ as } n \to \infty,$$

where δ_{G_n} is Bayes rule with respect to G_n, and $C = \{\eta : |\eta| < 1\}$.

Proof: Suppose the conditions of the theorem are satisfied but δ is inadmissible. Then we can find another estimator, say δ', whose risk doesn't uniformly exceed the risk of δ, i.e. $R(\eta, \delta') \leq R(\eta, \delta)$. At the same time δ' must differ from δ on the set of nonzero measure, i.e.

$$\int (\delta'(x,v) - \delta(x,v))^2\,\mathrm{d}x\,\mathrm{d}v > 0.$$

Define $\delta''(x,v) = \frac{1}{2}(\delta(x,v) - \delta'(x,v))$. Then,

$$\int (\delta'(x,v) - \delta(x,v))^2 \phi_{\eta,1}(x) f_\nu(v)\,\mathrm{d}x\,\mathrm{d}s = \mathsf{E}(\delta' - \delta)^2 > 0,$$

because $\phi_{\eta,1}(x)f_\nu(v) > 0$. However,

$$\begin{aligned}\mathsf{E}(\delta' - \delta)^2 &= \mathsf{E}((\eta - \delta) - (\eta - \delta'))^2 \\ &= \mathsf{E}(\eta - \delta)^2 + \mathsf{E}(\eta - \delta')^2 - 2\mathsf{E}(\eta - \delta)(\eta - \delta') > 0,\end{aligned}$$

and therefore

$$2\mathsf{E}(\eta - \delta)(\eta - \delta') < \mathsf{E}(\eta - \delta)^2 + \mathsf{E}(\eta - \delta')^2. \tag{4.11}$$

By inequality (4.11) we see that

$$\begin{aligned}R(\eta, \delta'') = \mathsf{E}(\eta - \delta'')^2 &= \frac{1}{4}\{\mathsf{E}(\eta - \delta)^2 + \mathsf{E}(\eta - \delta')^2 + 2\mathsf{E}(\eta - \delta)(\eta - \delta')\} \\ &< \frac{1}{2}(\mathsf{E}(\eta - \delta)^2 + \mathsf{E}(\eta - \delta')^2) = \frac{1}{2}(R(\eta, \delta) + R(\eta, \delta')) \leq R(\eta, \delta).\end{aligned}$$

Since $R(\eta, \delta)$ and $R(\eta, \delta'')$ are both continuous (by Lemma 1), there exists an $\varepsilon > 0$ such that $R(\eta, \delta'') < R(\eta, \delta) - \varepsilon$ for $|\eta| < 1$. Hence,

$$\int_C (R(\eta, \delta) - R(\eta, \delta''))G(d\eta) \geq \int_C \varepsilon G(d\eta) \geq \varepsilon, \tag{4.12}$$

and also

$$\varepsilon \leq r(G_n, \delta) - r(G_n, \delta'') \leq r(G_n, \delta) - r(G_n, \delta_{G_n}). \tag{4.13}$$

Inequality (4.13) contradicts (4.11) and thus proves the theorem. ||

Theorem 3 The usual estimator for η, $t(x, \Lambda, s) = x$ is

a. unbiased,

b. has constant risk equal 1,

c. admissible.

Proof: Clauses 1 and 2 are straightforward. The proof of 3 is based on Theorem 5.6.1 from Brown (1971) and in fact is nothing more than an extension of Stein's sufficient condition of admissibility (see Stein (1955)). This condition is formulated by us as Theorem 2. To apply Theorem 2 we need to check that there exists a sequence of nonnegative measures (generalized priors) G_i such that

a) $G_n(C) \geq 1$.

b) $r(G_n, \delta) - r(G_n, \delta_{G_n}) \to 0$, as $n \to \infty$.

Now let g_n be the rescaled normal density

$$\pi_n(\theta) = \exp(-\frac{\theta^2}{2n}).$$

Then claim a) follows because $G_n(C) = \sqrt{2\pi n}\ \mathsf{Erf}(\frac{1}{\sqrt{2n}})$, where $\mathsf{Erf}(\frac{y}{\sqrt{2}}) = 2\Phi(y) - 1$, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, and because $G_{n+1}(C) \geq G_n(C)$ and $G_1(C) \approx 1.71$. Then straightforward Bayes calculations give us

$$r(G_n, \delta) = \sqrt{2\pi}\sqrt{n}, \quad r(G_n, \delta_{G_n}) = \sqrt{2\pi}\frac{\sqrt{n^3}}{n+1},$$

that satisfy both a) and b). Therefore the usual estimator is admissible. ||

Theorem 4 The estimator $t^{(1)} = x/(1+c)$ is admissible for any $c > 0$.

Proof: The modified risk function (4.4) in this case boils down to the risk function for known σ and therefore the proof from Magnus (2000, Theorem A.2), can be applied without changes. That is: for $0 < c < \infty$ the normal Bayes estimators are Bayes with respect to prior $\pi(\eta) \sim \mathrm{N}(0, 1/c)$ and therefore admissible. For $c = \infty$, we have $\Lambda = 0$ and $R(0,0) = 0$, which implies admissibility. ||

Note, that $\mathrm{BIAS}(\eta, \Lambda) = \mathsf{E}_\eta \mathsf{E}_\nu(\Lambda(x\sqrt{\nu}/w)x - \eta)$ is an antisymmetric function, and that the risk $R(\eta, \Lambda)$ is a symmetric function with respect to η. This allows us to consider only positive η when investigating properties of the estimator.

Theorem 5 Suppose Λ is $R1$-regular function. Then $R(\eta, \Lambda)$ is bounded if there is a K, $0 \leq K \leq \infty$ such that $|\zeta(x)| \leq K$ for all x, where $\zeta(x) = (1 - \Lambda(x))x$.

Proof: Change of variable $x = u + \eta$, $u \sim \mathrm{N}(0,1)$ leads to

$$\begin{aligned} R(\eta, \Lambda) &= \mathsf{E}_0 \mathsf{E}_\nu(u - v\varepsilon(x\sqrt{\nu}/v))^2 \\ &\leq 2\mathsf{E}_0 \mathsf{E}_\nu(u^2 + v^2\varepsilon(x\sqrt{\nu}/v)^2) \leq 2(1 + K^2 A), \end{aligned}$$

where A is a second moment of the χ-distribution, that is a finite number. ||

The following result is useful when investigating properties of the WALS estimator in the auxiliary problem with unknown σ.

Theorem 6 Consider the estimation problem with known σ and suppose that the one-parametric class $\mathcal{L}^{(a)} = \{\Lambda_c : 0 \geq c \geq \infty\}$ consists of $R1$-regular estimators that are $\mathcal{L}^{(a)}$-admissible. Assume also that

$$\Lambda_c(\frac{x}{s}) = \Lambda_{cs}(x), \tag{4.14}$$

for any $s \geq 0$. Then the estimator $\Lambda_c(\frac{x}{s})x$ with risk function $R(\eta, \Lambda_c) = \mathsf{E}_s R_s(\eta, \Lambda_c(\frac{\cdot}{s}))$ is $\mathcal{L}^{(a)}$-admissible.

Proof: Suppose that the opposite is true and that the estimator $\Lambda_c(\frac{x}{s})x$ is not admissible for some $c \geq 0$. Then we can find c_o such that $R(\eta, \Lambda_{c_o}) \leq R(\eta, \Lambda_c)$, and such that for at least one η, say η_o, this inequality is strict. Then,

$$\mathsf{E}_s(R_s(\eta_o, \Lambda_{c_o}, s) - R_s(\eta_o, \Lambda_c, s)) < 0,$$

so there is an $s > 0$ such that $R_s(\eta_o, \Lambda_{c_o}, s) < R_s(\eta_o, \Lambda_c, s)$. Therefore $R(\eta_o, \Lambda_{c_o s}) < R(\eta_o, \Lambda_{cs})$ by property (4.14). Hence inadmissibility follows for $\Lambda_{c'}(x)x$, $c' = cs$. This contradicts the definition of $\mathcal{L}^{(a)}$. ||

The following two propositions are direct consequences of Theorem 6.

Proposition 7 The pretest estimator (4.7) in the auxiliary problem with unknown variance is $\mathcal{L}^{(1)}$-admissible.

Proposition 8 The Burr estimator (4.9) in the auxiliary problem with unknown variance is $\mathcal{L}^{(2)}$-admissible.

Theorem 9 The Laplace estimator (4.8) in the auxiliary problem with unknown variance is admissible.

Proof: The main idea of proof is close to the proof of Theorem 6. Suppose that the Laplace estimator is inadmissible. Then there exists an estimator $\delta(x, s)$ such that $R(\eta, \delta) \leq R(\eta, \Lambda^{(4)})$ for any η and there is at least one point η_o where this inequality is strict. Then,

$$\mathsf{E}_s(R_s(\eta_o, \delta) - R_s(\eta_o, \Lambda^{(4)})) < 0,$$

and therefore $R_s(\eta, \delta(\cdot, s)) - R_s(\eta, \Lambda^{(4)}(\frac{\cdot}{s})) < 0$ for some $s > 0$. This implies inadmissibility for the estimator

$$\Lambda^{(4)}_{c,s}(x) = \Lambda^{(4)}(\frac{x}{s}) \tag{4.15}$$

in the problem with known σ. However, the estimator (4.15) is a Bayes estimator by Lemma 10 and therefore admissible. ||

Lemma 10 The estimator (4.15) is a Bayes estimator with prior $\pi(\eta, c_1) = \frac{c_1}{2} e^{-c_1|\eta|}$, $c_1 = c/s$ and $x|\eta \sim \mathrm{N}(\eta, s^2)$.

Proof: Performing Bayes calculations we get $t(x; c_1, s) = x - h_1(c_1, x)c_1 s^2$, where

$$h_1(c_1, x) = \frac{1 - e^{2c_1 x} d_1(c_1, x)}{1 + e^{2c_1 x} d_1(c_1, x)}, \quad d_1(c_1, x) = \frac{\Phi(\frac{-x - c_1 s^2}{s})}{\Phi(\frac{x + c_1 s^2}{s})}.$$

But

$$t(x; c_1, s) = x(1 - \frac{h_1(c_1, x)}{x/s} c_1 s) = x\Lambda^{(4)}_{c_1 s}(x/s) = x\Lambda^{(4)}_c(x/s),$$

and the result follows. ||

Theorem 11

$$\inf_{\lambda \in \mathcal{L}^{(0)}} R(\eta, \lambda) = \frac{\eta^2}{1 + \eta^2},$$

where $\mathcal{L}^{(0)}$ includes all $R1$-regular Λ-s.

Proof: We need to check

a) for any $\Lambda \in \mathcal{L}^{(0)}$ the inequality $R(\eta, \Lambda) \geq \frac{\eta^2}{1+\eta^2}$ holds.

b) There exist at least one $\Lambda' \in \mathcal{L}^{(0)}$ such that $R(\eta, \Lambda') = \frac{\eta^2}{1+\eta^2}$.

To prove a) consider the definition of $R(\eta, \Lambda)$, that is

$$R(\eta, \Lambda) = \mathsf{E}_\eta \mathsf{E}_\nu (\Lambda(x\sqrt{\nu}/v)x - \eta)^2 = \mathsf{E}_\nu R_v(\eta, \Lambda).$$

Due to (4.5), $R_v(\eta, \Lambda)$ can be interpreted as the risk function for the problem with known σ but with modified Λ, i.e. $\Lambda_v = \Lambda(x\sqrt{\nu}/v)$. However, *any* $R_v(\eta, \Lambda)$ must satisfy $R_v(\eta, \Lambda) \geq \frac{\eta^2}{1+\eta^2}$ by Theorem A.7 in Magnus(2000). Therefore,

$$R(\eta, \Lambda) = \mathsf{E}_\nu R_v(\eta, \Lambda) \geq \mathsf{E}_\nu \frac{\eta^2}{1+\eta^2} = \frac{\eta^2}{1+\eta^2}.$$

To prove b) just consider $\Lambda = \frac{1}{1+1/\eta^2}$.

||

Chapter 5

Conclusions

Pretesting phenomena have been a difficult aspect in econometrics for at least sixty years. This thesis contributes to the extensive literature on the topic in three aspects. The theoretical contribution consists of the generalization of the 'equivalence theorem', a powerful tool for investigation of the mean squared error performance of weighted-average least-squares (WALS) estimators. Various applications of the equivalence theorem help in understanding theoretical and empirical aspects of pretest estimation and prediction. The empirical contribution consists in applying the theoretical tools to an empirically relevant financial example. We check — for the first time in the literature — how large the actual distortion is of the accuracy of forecasting schemes, used in a specific application. The final contribution proposes a way to improve the properties of the pretest estimator, and investigates a feasible version of the 'neutral Laplace' estimator.

In the following sections we consider these contributions in more detail, and also discuss some possible extensions and open questions.

5.1 On the harm that pretesting does

In Chapter 2 we derive the bias, variance, and mean squared error of the pretest estimator under very general assumptions. The 'equivalence theorem' of Magnus and Durbin (1999) is generalized to an arbitrary number of auxiliary regressors and an arbitrary number of pretestings. The new equivalence theorem allows us to investigate the mean squared error, bias,

and variance for a number of model selection procedures that are relevant in practice.

The theory is then applied to the following major problem. It has been known for a long time that ignoring pretesting could lead to a distortion of the accuracy of subsequent estimation. However, it was not known how large the actual distortion is. We compare the exact moments of the pretest estimator with the moments of the OLS estimators, that are routinely reported in applied studies. We conclude that underreporting is a very serious problem and that not reporting the correct pretest bias and variance can lead to very misleading results. Even in the case of just one auxiliary regressor, the variance of the estimator can be misreported by a factor of 7.5. The pretest bias appears to be less of a problem than the pretest variance.

For the case of two auxiliary regressors we consider two widely accepted model selection procedures: specific-to-general and general-to-specific. We find that for the specific-to-general model selection procedure, underreporting can in fact be unbounded, that is, the reported (OLS) variance may have absolutely nothing to do the with the actual (pretest) variance. We also find that there is a large difference in performance between these two model selection procedures. The influence of the selection procedure is greater when the correlation between the $\boldsymbol{z}$ variables (measured by $\boldsymbol{Z'MZ}$) is high than when it is low. If we can choose the auxiliary regressors in such a way that they are 'orthogonal' (that is, $\boldsymbol{Z'MZ}$ equals the identity matrix), then all pretest procedures are the same, and hence the sampling properties of the estimators do not depend on the model selection procedure. For this 'orthogonal' case we study the question: how fast does the underreporting of the variance grow when the number of auxiliary regressors grows. We find that as the number of auxiliary regressors grows, the average underreporting grows, but not very fast. We also find that our results are robust with respect to regression model misspecification. Moreover, the underreporting does not disappear even asymptotically when the sample size grows, at least if the level of the preliminary test is kept constant.

Since all practically important model selection procedures are a function of $\boldsymbol{My}$, our equivalence theorem applies to these cases too. Therefore, future work should investigate other model selection procedures not yet covered by our research.

5.2 Forecast accuracy after pretesting

While Chapter 2 deals with estimation, in Chapter 3 we investigate how pretesting affects the accuracy of the one-step-ahead regression forecast, and apply the results to the stock market. The most common situation in empirical analysis is when forecasting is preceded by preliminary model selection. We introduce the WALS forecast procedure, which generalizes the procedure of forecasting using a regression model selected by preliminary tests. Our definition of the WALS forecast not only includes the majority of classical model selection procedures as special cases, but also allows for generalizations. For example, 'smooth' combinations of model selection procedures and application of the Bayesian approach become possible. Then, we generalize the equivalence theorem of Chapter 2, and obtain the unconditional moments of the one-step-ahead WALS forecast.

The theoretical results are then applied to the stock market, a case considered in Pesaran and Timmermann (1994). For the model selection procedure described by the authors we calculate the unconditional moments and evaluate the accuracy of their forecasts. We find that the model selection procedure seriously affects the standard errors of the forecasts, and thus the accuracy reported in their article is noticeably overestimated. Granger and Pesaran (2000) extend the standard regression approach to forecasting by introducing the probability forecast. For the stock market case this probability forecast is built on the regression forecast and therefore suffers the same drawback: it's actual variability is underestimated. Our study shows that the underreporting of the variance in this case is even larger than for the point forecast error.

Next, we propose several ways to improve the accuracy of the forecast, in particular by orthogonalizing the auxiliary regressors. In addition we consider a set of problems that arise due to estimation of the moments of WALS estimator: the natural estimators of these moments are in fact biased and inconsistent. More conservative interval estimators of the accuracy are proposed.

Thus, the chapter not only investigates the effects of pretesting on the forecast accuracy, but also shows in a real-life example that not reporting the actual moments significantly distorts the accuracy of the forecasts. Of course, more work is needed to find better estimators for the pretest bias, variance, and mean squared error.

5.3 Estimation of the mean of a univariate normal distribution when variance is not known

Chapter 4 contains an extensive treatment of the neutral Laplace WALS estimator. The Laplace WALS estimator was introduced in Magnus (2002) under the assumption that the variance of the regression innovations is known. He reduced the problem of optimal WALS regression estimation to the problem of optimal estimation of the unknown location parameter for the normal distribution with unit variance, when only one observation is available. The optimal solution then obtained is based on a (Bayesian) WALS estimator with a 'neutral' Laplace prior.

We now investigate the properties of the WALS estimator in a regression problem in the obviously more realistic case where the variance is not known. We propose to estimate the unknown variance by the least-squares estimator in the unrestricted model. Then we investigate the properties of this feasible version of the Laplace estimator.

We show that in the new set-up the Laplace estimator is admissible, and its risk and regret change only marginally, compared to the known-variance case. Therefore, no modification of the tuning parameter is necessary, and exactly the same λ-function can be used in both situations, with or without the assumption of known variance.

We also investigate the performance of a competing minimax regret estimator, the so-called Burr estimator. We find that choosing the tuning parameter according to the minimax regret principle has only a minor influence on the performance. Finally, we check the performance of the feasible neutral Laplace estimator in a regression problem and compare its performance with the performance of the usual pretest estimator. We find that the Laplace estimator performs better over the practically important range of the parameter. The superiority of the Laplace estimator is more pronounced for small sample sizes. On the basis of our study, we recommend the neutral Laplace estimator to be used at practice.

Chapter 6

Nederlandse samenvatting

Het voorbereidend toetsen is al een moeilijk aspect in de econometrie gedurende de afgelopen zestig jaar. Dit proefschrift draagt bij aan de uitgebreide literatuur over dit onderwerp in drie aspecten. De theoretische bijdrage bestaat uit de generalisatie van de 'equivalentie stelling', een krachtig hulpmiddel voor onderzoek naar de gemiddelde kwadratische fout performance van de gewogen-gemiddelde kleinste kwadraten (*WALS*) schatters. Verscheidene toepassingen van de equivalentie stelling helpen ons theoretische en empirische aspecten van pretest schatten en voorspellen te begrijpen. De empirische bijdrage bestaat uit het toepassen van de theoretische hulpmiddelen op een empirisch relevant voorbeeld uit de financiering. We testen - voor het eerst in de literatuur - hoe groot de eigenlijke verstoring van de nauwkeurigheid van voorspel concepten is, gebruik makend van een specifieke toepassing. De laatste bijdrage is het voorstel ter verbetering van de eigenschappen van de pretest schatter en onderzoekt een uitvoerbare versie van de 'neutrale Laplace' schatter.

In de volgende paragrafen beschouwen we deze bijdragen nauwkeuriger en bediscussieren mogelijke uitbreidingen en open vragen.

6.1 Over de schade die voorbereidende toetsen aanricht

In Hoofdstuk 2 leiden we de onzuiverheid, variantie, en gemiddelde kwadratische fout van de pretest schatter af onder erg algemene aannames. De 'equiv-

alentie stelling' van Magnus en Durbin (1999) is gegeneraliseerd naar een willekeurig aantal hulp regressoren en een willekeurig aantal pretests. De nieuwe equivalentie stelling laat ons toe om de gemiddelde kwadratische fout, onzuiverheid, en variantie voor een aantal model selectie procedures die relevant zijn in de praktijk te onderzoeken.

De theorie wordt dan toegepast op het volgende belangrijke probleem. Het is reeds lange tijd bekend dat het negeren van voorbereidende toetsen kan leiden tot een verstoring van de nauwkeurigheid van daaropvolgend schatten. Het was echter niet bekend hoe groot deze verstoring is. We vergelijken de exacte momenten van de pretest schatter met de momenten van de *OLS* schatters die gewoontegetrouw gerapporteerd worden in toegepaste studies. We concluderen dat *underreporting* een erg ernstig probleem is en dat het niet rapporteren van de pretest onzuiverheid en variantie kan leiden tot erg misleidende resultaten. Zelfs in het geval van slechts een hulp regressor, de variantie van de schatter kan met een factor 7.5 verkeerd gerapporteerd worden. De pretest onzuiverheid blijkt een minder groot probleem te zijn dan de pretest variantie.

Voor het geval van twee hulp regressoren beschouwen we twee gangbare model selectie procedures: specifiek-naar-algemeen en algemeen-naar-specifiek. We vinden dat voor de specifiek-naar-algemeen model selectie procedure, *underreporting* in feite onbegrensd kan zijn, wat betekent dat de gerapporteerde (*OLS*) variantie niets te maken heeft met de eigenlijke (pretest) variantie. We vinden ook dat er een groot verschil in performance is tussen deze twee model selectie procedures. De invloed van de model selectie procedure is groter wanneer de correlatie tussen de $\boldsymbol{z}$ variabelen (gemeten door $\boldsymbol{Z'MZ}$) hoog is dan wanneer deze laag is. Als we de hulp regressoren zo kunnen kiezen dat ze 'orthogonaal' zijn (dat betekent, $\boldsymbol{Z'MZ}$ is gelijk aan de identiteits matrix), dan zijn alle pretest procedures hetzelfde, en dus zijn de steekproef eigenschappen onafhankelijk van de model selectie procedure. Voor dit 'orthogonale' geval bestuderen we de vraag: Hoe snel groeit de *underreporting* van de variantie als het aantal hulp regressoren toeneemt? We vinden dat als het aantal hulp regressoren toeneemt de gemiddelde *underreporting* groeit, maar niet erg snel. We vinden ook dat onze resultaten robuust zijn met betrekking tot misspecificatie van het regressie model. Bovendien verdwijnt de *underreporting* zelfs asymptotisch niet als de omvang van de steekproef toeneemt, tenminste als de significantieniveau van de voorbereidende toetsen constant gehouden wordt.

Aangezien alle praktisch belangrijke model selectie procedures een functie zijn van $\boldsymbol{My}$, geldt onze equivalentie stelling ook in deze gevallen. Derhalve zou toekomstig onderzoek zich moeten richten op model selectie procedures die nog niet door ons onderzoek omvat worden.

6.2 Voorspelnauwkeurigheid na voorbereidend toetsen

Terwijl in Hoofdstuk 2 schatten behandeld wordt, onderzoeken we in Hoofdstuk 3 hoe voorbereidend toetsen de nauwkeurigheid van de een-periode-vooruit regressie voorspelling beinvloed, en passen deze resultaten toe op de aandelen markt. The meest gebruikelijke situatie in empirische analyses is dat voorspellen door voorbereidende model selectie wordt vooraf gegaan. We introduceren de *WALS* voorspellingsprocedure, die een veralgemenisering is van de voorspellingsprocedure die gebruik maakt van een regressie model geselecteerd door voorbereidende toetsen. Onze definitie van de *WALS* voorspeller omvat niet alleen de meerderheid van de klassieke model selectie procedures als een speciaal geval, maar laat ook veralgemeniseringen toe. Bijvoorbeeld, [smooth] combinaties van model selectie procedures en toepassingen van de Bayesiaanse aanpak zijn mogelijk. Verder generaliseren we de equivalentie stelling van Hoofdstuk 2, en verkrijgen de onconditionele momenten van de een-periode-vooruit *WALS* voorspeller.

De theoretische resultaten worden toegepast op de aandelenmarkt, een geval dat in Pesaran en Timmermann (1994) beschouwd wordt. We berekenen de onconditionele momenten voor de model selectie procedure die beschreven wordt door deze auteurs, en evalueren de nauwkeurigheid van hun voorspellingen. We vinden dat de model selectie procedure een aanzienlijke invloed heeft op de standaard fouten van de voorspellingen, en dus wordt de nauwkeurigheid die gerapporteerd wordt in hun artikel danig overschat. Granger en Pesaran (2000) breiden de standaard regressie aanpak uit naar voorspellen door de introductie van de kansvoorspelling. Voor de aandelenmarkt toepassing is deze kansvoorspelling opgebouwd uit regressie voorspellingen en derhalve lijdt deze methode aan hetzelfde bezwaar: de eigenlijke variabiliteit wordt onderschat. Onze studie laat zien dat het onderschatten van de variantie in dit geval nog groter is dan in het geval van de fout behorende bij de punt voorspelling.

Vervolgens stellen we enkele methodes voor om de nauwkeurigheid van de schattingen te verbeteren, in het bijzonder door de hulpvariabelen te orthogonaliseren. Bovendien beschouwen we een verzameling problemen die ontstaat door het schatten van de momenten van de *WALS* schatter: de natuurlijke schatters voor deze momenten zijn onzuiver en inconsistent. Er worden meer conservatieve interval schatters voor de nauwkeurigheid aangedragen.

Het hoofdstuk onderzoekt dus niet enkel de effecten van voorbereidend toetsen op de voorspelnauwkeurigheid, maar laat ook zien dat in een praktisch voorbeeld de nauwkeurigheid van de voorspellingen significant verstoord wordt indien de eigenlijke momenten niet gerapporteerd worden. Natuurlijk is er nog meer werk vereist om betere schatters voor de pretest onzuiverheid, variantie, en gemiddelde kwadratische fout te vinden.

6.3 Schatten van de verwachting van een univariate normale verdeling met onbekende variantie

Hoofdstuk 4 bevat een uitgebreide uiteenzetting van de neutrale Laplace *WALS* schatter. De Laplace *WALS* schatter is geintroduceerd in Magnus (2002) onder de veronderstelling dat de variantie van de regressie innovaties bekend is. Hij reduceert het probleem van optimale *WALS* regressie schatting tot het probleem van optimale schatting van de onbekende locatie parameter van de normale verdeling met een variantie van een, met slechts een beschikbare waarneming. De hierbij verkregen optimale oplossing is gebaseerd op een (Bayesiaanse) *WALS* schatter met een 'neutrale' Laplace prior.

We onderzoeken nu de eigenschappen van de *WALS* schatter in een regressie probleem in het duidelijk meer realistische geval wanneer de variantie onbekend is. We stellen voor om de onbekende variantie te schatten met de kleinste-kwadraten schatter in het ongerestringeerde model. Vervolgens onderzoeken we de eigenschappen van de uitvoerbare versie van de Laplace estimator.

We laten zien dat in de nieuwe opzet de Laplace schatter toegelaten is, en het risico and *regret* slechts marginaal veranderen in vergelijking met het geval waarin de variantie bekend is. Derhalve is een verandering van de afstemmings parameters niet nodig en precies dezelfde λ-functie kan gebruikt

worden met of zonder de aanname van bekende variantie.

We onderzoeken ook de performance van een rivaliserende *minimax regret* schatter, de zogenaamde Burr schatter. We vinden dat het kiezen van de afstemmingsparameter volgens het minimax spijt principe slechts een beperkte invloed heeft op de performance. Tenslotte controleren we de performance van de uitvoerbare neutrale Laplace schatter in een regressie probleem en vergelijken deze met de performance van de gebruikelijke pretest schatter. We vinden dat de Laplace schatter beter presteert over de praktisch belangrijke bereik van de parameter. De superioriteit van de Laplace schatter is meer uitgesproken voor kleine steekproeven. Op basis van onze studie bevelen we de neutrale Laplace schatter aan voor praktisch gebruik.

Bibliography

Abadir, K.M. and J.R. Magnus (2002). Notation in econometrics: a proposal for a standard. *The Econometrics Journal*, 5, 76–90.

Anderson, T.W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics*, 33, 255–265.

Balvers, R.J., T.F. Cosimano and B. McDonald (1990). Predicting stock returns in an efficient market. *Journal of Finance*, 55, 1109–1128.

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics*, 15, 190–204.

Berger, J.O. (1976). Admissibility results for generalised bayes estimators of coordinates of a location vector. *The Annals of Statistics*, 4, 334–356.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of American Statistical Assosiation*, 37, 325–335.

Black, F., M. Jensen and M. Scholes (1972). The capital asset pricing model: Some empirical tests. In M. Jensen (ed.), *Studies in the Theory of Capital Markets*, Praeger, New York.

Blyth, C.R. (1951). On minimax statistical decision procedures and their admissibility. *Annals of Mathematical Statistics*, 22, 22–42.

Brook, R.J. (1976). On the use of a regret function to set significance points in prior tests of estimation. *Journal of American Statistical Assosiation*, 71, 126–131.

Brown, L.D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Annals of Mathematical Statistics*, 42, 855–903.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society*, Series A, 158, 419–466.

Chen, N.F., R. Roll and S.A. Ross (1986). Economic forces and the stock market. *Journal of Business*, 59, 382–403.

Cheng, T.C.E., Y.K. Lo and K.W. Ma (1990). Forecasting stock price index by multiple regression. *Managerial Finance*, 16, 27–32.

Danilov, D. (2002). Estimation of the mean of a univariate normal distribution when the variance is not known. CentER Discussion Paper, Tilburg University, The Netherlands. [Chapter 4 of this thesis]

Danilov, D. and J.R. Magnus (2001). On the harm that pretesting does. CentER Discussion Paper, Tilburg University, The Netherlands. [Chapter 2 of this thesis]

Danilov, D. and J.R. Magnus (2002). Forecast accuracy after pretesting with an application to the stock market. CentER Discussion Paper, Tilburg University, The Netherlands. [Chapter 3 of this thesis]

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–265.

Droge B. and T. Georg (1995). On selecting the smoothing parameter of least squares regression estimates using the minimax regret approach. *Statistics and Decisions*, 13, 1–20.

Fama, E. and K.R. French (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25, 23–49.

Fama, E. and K.R. French (1992). The cross-section of expected stock returns. *Journal of Finance*, 47, 427–465.

Fama, E. and J. MacBeth (1973). Risk, return and equilibrium: empirical tests. *Journal of Political Economy*, 71, 607–636.

Farrell, R.H. (1968). On a necessary and sufficient condition for admissibility of estimators when strictly convex loss is used. *The Annals of Mathematical Statistics*, 38, 23–28.

Feldstein, M.S. (1973). Multicollinearity and the mean square error of alternative estimators. *Econometrica*, 41, 337–346.

French, K.R., G.W. Schwert, and R. Stambaugh (1987). Expected stock returns and volatility. *Journal of Financial Economics*, 19, 3–30.

George, E.I. (2000). The variable selection problem. *Journal of American Statistical Assosiation*, 95, 1304–1308.

Giles, D.E.A. (2000). Preliminary-test and Bayes estimation of a location parameter under "reflected normal" loss. Econometrics Working Papers N. 0004, Department of Economics, University of Victoria.

Giles, J. (1991). Pre-testing for linear restrictions in a regression model with spherically symmetric disturbances. *Journal of Econometrics*, 50, 377–398.

Giles, J.A. and D.E.A. Giles (1993). Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys*, 7, 145–197.

Giles, J.A. and V. Srivastava (1993). The exact distribution of a least-squares regression coefficient after a preliminary t-test. *Statistics and Probability Letters*, 16, 59–64.

Goldberger, A. (1981). Linear regression after selection. *Journal of Econometrics*, 15, 357–366.

Granger, C.W.J. and M.H. Pesaran (2000). Economic and Statistical Measures of Forecast Accuracy. *Journal of forecasting*, 19, 537–560

Hendry, D.F. (2001). Achievements and challenges in econometric methodology. *Journal of Econometrics*, 100, 7–10.

Hoover, K.D. and S.J. Perez (1999). Data mining reconsidered; encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, 2, 1–25.

Huntsberger, D.V. (1955). A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics*, 26, 734–743.

James, W. and C. Stein (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* , 1, (Berkeley, University of California Press), 361–379.

Judge, G.G. and M.E. Bock (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.

Judge, G.G. and M.E. Bock (1983). Biased estimation. In Z. Griliches, and M.D. Intriligator (eds.), *Handbook of Econometrics*, Vol. I, Chapter 10, Amsterdam: North-Holland.

Karlin, S. (1958). Admissibility for estimator with quadratic loss. *Annals of Mathematical Statistics*, 29, 406–436.

Leeb, H. and B.M. Pötscher (2000). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. Discussion paper, Department of Economics, Universität Wien.

Lo, A.W. and A.C. MacKinlay (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3, 431–467.

Lovell, M.C. (1983). Data mining. *The Review of Economics and Statistics*, 65, 1–12.

Lucas, R.E., Jr. (1976). Econometric policy evaluation: a critique. In K. Brunner and A.H. Meltzer (eds.), *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy, Vol. 1, Amsterdam: North-Holland, 19–46.

Magnus, J.R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications*, 44, 293–308.

Magnus, J.R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *The Econometrics Journal*, 5, 225–236.

Magnus, J.R. and J. Durbin (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica*, 67, 639–643.

Miller, A.J. (1990). *Subset Selection in Regression.* London: Chapman and Hall.

Mittelhammer, R.C. (1984). Restricted least squares, pre-test, OLS and Stein rule estimators: Risk comparisons under model misspecification. *Journal of Econometrics*, 25, 151–164.

Mosteller, F. (1948). On pooling data. *Journal of the American Statistical Association*, 43, 231–242.

Ohtani, K., D.E.A. Giles and J.A. Giles (1997). The exact risk performance of a pre-test estimator in a heteroskedastic linear restriction model under the balanced loss function. *Econometric reviews*, 16, 119–130.

Pesaran, M.H. and A. Timmermann (1994). Forecasting stock returns. An examination of stock market trading in the presence of transaction costs. *Journal of Forecasting*, 13, 335–367.

Pesaran, M.H. and A. Timmermann (1995). Predictability of stock returns: robustness and economic significance. *Journal of Finance*, 50, 1201–1228.

Pötscher, B.M. (1983). Order estimation in ARMA-models by Lagrangian multiplier tests. *The Annals of Statistics*, 11, 872–885.

Pötscher, B.M. (1991). Effects of model selection on inference. *Econometric Theory*, 7, 163–185.

Pötscher, B.M. and A.J. Novak (1998). The distribution of estimators after model selection: large and small sample results. *Journal of Statistical Computation and Simulation*, 60, 19–56.

Roehrig, C.S. (1984). Optimal critical regions for pre-test estimators using a Bayes risk criterion. *Journal of Econometrics*, 25, 3–14.

Rozeff, M.S. (1984). Dividend yields are equity risk premiums. *Journal of Portfolio Management*, 10, 68–75.

Sawa, T. and T. Hiromatsu (1973). Minimax regret significance points for a preliminary test in regression analysis. *Econometrica*, 41, 1093–1101.

Sclove, S.L., C. Morris, and R. Radhakrishnan (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, 43, 1481–1490.

Sen, P.K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics*, 7, 1019–1033.

Stein, C. (1955). A necessary and sufficient condition for admissibility. *Annals of Mathematical Statistics*, 26, 518–522.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, (Berkeley, University of California Press), 197–206.

Thomson, M. and P. Schmidt (1982). A note on the comparison of the mean square error of inequality constrained least squares and other related estimators. *The Review of Economics and Statistics*, 64, 174–176.

Toyoda, T. and T. Wallace (1976). Optimal critical values for pre-testing in regression. *Econometrica*, 44, 365–375.

Wallace, T.D. and V.G. Ashtar (1972). Sequential methods in model construction. *The Review of Economics and Statistics*, 54, 172–178.

Wan, A.T.K. (1996). Estimating the error variance after a pre-test for an inequality restriction on the coefficients. *Journal of Statistical Planning and Inference*, 52, 197–213.

West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64, 1067–1084.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68, 1097–1126.

Wolfowitz, J. (1950). Minimax estimates of the mean of a normal distribution with known variance. *Annals of Mathematical Statistics*, 21, 218–230.

Zaman, A. (1984). Avoiding model selection by the use of shrinkage techniques. *Journal of Econometrics*, 25, 73–85.

Zhang, P. (1992). On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, 87, 732–737.

Center for Economic Research, Tilburg University, The Netherlands
Dissertation Series

No.	Author	Title	Published
1	P.J.J. Herings	Static and Dynamic Aspects of General Disequilibrium Theory; ISBN 90 5668 001 3	June 1995
2[1]	Erwin van der Krabben	Urban Dynamics: A Real Estate Perspective - An institutional analysis of the production of the built environment; ISBN 90 5170 390 2	August 1995
3	Arjan Lejour	Integrating or Desintegrating Welfare States? - a qualitative study to the consequences of economic integration on social insurance; ISBN 90 5668 003 X	September 1995
4	Bas J.M. Werker	Statistical Methods in Financial Econometrics; ISBN 90 5668 002 1	September 1995
5	Rudy Douven	Policy Coordination and Convergence in the EU; ISBN 90 5668 004 8	September 1995
6	Arie J.T.M. Weeren	Coordination in Hierarchical Control; ISBN 90 5668 006 4	September 1995
7	Herbert Hamers	Sequencing and Delivery Situations: a Game Theoretic Approach; ISBN 90 5668 005 6	September 1995
8	Annemarie ter Veer	Strategic Decision Making in Politics; ISBN 90 5668 007 2	October 1995
9	Zaifu Yang	Simplicial Fixed Point Algorithms and Applications; ISBN 90 5668 008 0	January 1996
10	William Verkooijen	Neural Networks in Economic Modelling - An Empirical Study; ISBN 90 5668 010 2	February 1996
11	Henny Romijn	Acquisition of Technological Capability in Small Firms in Developing Countries; ISBN 90 5668 009 9	March 1996
12	W.B. van den Hout	The Power-Series Algorithm - A Numerical Approach to Markov Processes; ISBN 90 5668 011 0	March 1996
13	Paul W.J. de Bijl	Essays in Industrial Organization and Management Strategy; ISBN 90 5668 012 9	April 1996

No.	Author	Title	Published
14	Martijn van de Ven	Intergenerational Redistribution in Representative Democracies; ISBN 90 5668 013 7	May 1996
15	Eline van der Heijden	Altruism, Fairness and Public Pensions: An Investigation of Survey and Experimental Data; ISBN 90 5668 014 5	May 1996
16	H.M. Webers	Competition in Spatial Location Models; ISBN 90 5668 015 3	June 1996
17	Jan Bouckaert	Essays in Competition with Product Differentiation and Bargaining in Markets; ISBN 90 5668 016 1	June 1996
18	Zafar Iqbal	Three-Gap Analysis of Structural Adjustment in Pakistan; ISBN 90 5668 017 X	September 1996
19	Jimmy Miller	A Treatise on Labour: A Matching-Model Analysis of Labour-Market Programmes; ISBN 90 5668 018 8	September 1996
20	Edwin van Dam	Graphs with Few Eigenvalues - An interplay between combinatorics and algebra; ISBN 90 5668 019 6	October 1996
21	Henk Oosterhout	Takeover Barriers: the good, the bad, and the ugly; ISBN 90 5668 020 X	November 1996
22	Jan Lemmen	Financial Integration in the European Union: Measurement and Determination; ISBN 90 5668 021 8	December 1996
23	Chris van Raalte	Market Formation and Market Selection; ISBN 90 5668 022 6	December 1996
24	Bas van Aarle	Essays on Monetary and Fiscal Policy Interaction: Applications to EMU and Eastern Europe; ISBN 90 5668 023 4	December 1996
25	Francis Y. Kumah	Common Stochastic Trends and Policy Shocks in the Open Economy: Empirical Essays in International Finance and Monetary Policy; ISBN 90 5668 024 2	May 1997
26	Erik Canton	Economic Growth and Business Cycles; ISBN 90 5668 025 0	September 1997

No.	Author	Title	Published
27	Jeroen Hoppenbrouwers	Conceptual Modeling and the Lexicon; ISBN 90 5668 027 7	October 1997
28	Paul Smit	Numerical Analysis of Eigenvalue Algorithms Based on Subspace Iterations; ISBN 90 5668 026 9	October 1997
29	Uri Gneezy	Essays in Behavioral Economics; ISBN 90 5668 028 5	October 1997
30	Erwin Charlier	Limited Dependent Variable Models for Panel Data; ISBN 90 5668 029 3	November 1997
31	Rob Euwals	Empirical Studies on Individual Labour Market Behaviour; ISBN 90 5668 030 7	December 1997
32	Anurag N. Banerjee	The Sensitivity of Estimates, Inferences, and Forecasts of Linear Models; ISBN 90 5668 031 5	December 1997
33	Frans A. de Roon	Essays on Testing for Spanning and on Modeling Futures Risk Premia; ISBN 90 5668 032 3	December 1997
34	Xiangzhu Han	Product Differentiation, Collusion and Standardization; ISBN 90 5668 033 1	January 1998
35	Marcel Das	On Income Expectations and Other Subjective Data: A Micro-Econometric Analysis; ISBN 90 5668 034 X	January 1998
36	Jeroen Suijs	Cooperative Decision Making in a Stochastic Environment; ISBN 90 5668 035 8	March 1998
37	Talitha Feenstra	Environmental Policy Instruments and International Rivalry: A Dynamic Analysis; ISBN 90 5668 036 6	May 1998
38	Jan Bouwens	The Use of Management Accounting Systems in Functionally Differentiated Organizations; ISBN 90 5668 037 4	June 1998
39	Stefan Hochguertel	Households' Portfolio Choices; ISBN 90 5668 038 2	June 1998
40	Henk van Houtum	The Development of Cross-Border Economic Relations; ISBN 90 5668 039 0	July 1998
41	Jorg Jansen	Service and Inventory Models subject to a Delay-Limit; ISBN 90 5668 040 4	September 1998

No.	Author	Title	Published
42	F.B.S.L.P. Janssen	Inventory Management Systems: control and information issues; ISBN 90 5668 041 2	September 1998
43	Henri L.F. de Groot	Economic Growth, Sectoral Structure and Unemployment; ISBN 90 5668 042 0	October 1998
44	Jenke R. ter Horst	Longitudinal Analysis of Mutual Fund Performance; ISBN 90 5668 043 9	November 1998
45	Marco Hoeberichts	The Design of Monetary Institutions; ISBN 90 5668 044 7	December 1998
46	Adriaan Kalwij	Household Consumption, Female Employment and Fertility Decisions: A microeconometric analysis; ISBN 90 5668 045 5	February 1999
47	Ursula Glunk	Realizing High Performance on Multiple Stakeholder Domains: A Resource-based Analysis of Professional Service Firms in the Netherlands and Germany; ISBN 90 5668 046 3	April 1999
48	Freek Vermeulen	Shifting Ground: Studies on the Intersection of Organizational Expansion, Internationalization, and Learning; ISBN 90 5668 047 1	February 1999
49	Haoran Pan	Competitive Pressures on Income Distribution in China; ISBN 90 5668 048 X	March 1999
50	Laurence van Lent	Incomplete Contracting Theory in Empirical Accounting Research; ISBN 90 5668 049 8	April 1999
51	Rob Aalbers	On the Implications of Thresholds for Economic Science and Environmental Policy; ISBN 90 5668 050 1	May 1999
52	Abe de Jong	An Empirical Analysis of Capital Structure Decisions in Dutch Firms; ISBN 90 5668 051 X	June 1999
53	Trea Aldershof	Female Labor Supply and Housing Decisions; ISBN 90 5668 052 8	June 1999
54	Jan Fidrmuc	The Political Economy of Reforms in Central and Eastern Europe; ISBN 90 5668 053 6	June 1999
55	Michael Kosfeld	Individual Decision-Making and Social Interaction; ISBN 90 5668 054 4	June 1999

No.	Author	Title	Published
56	Aldo de Moor	Empowering Communities - A method for the legitimate user-driven specification of network information systems; ISBN 90 5668 055 2	October 1999
57	Yohane A. Khamfula	Essays on Exchange Rate Policy in Developing Countries; ISBN 90 5668 056 0	September 1999
58	Maurice Koster	Cost Sharing in Production Situations and Network Exploitation; ISBN 90 5668 057 9	December 1999
59	Miguel Rosellón Cifuentes	Essays on Financial Policy, Liquidation Values and Product Markets; ISBN 90 5668 058 7	December 1999
60	Sharon Schalk	Equilibrium Theory: A salient approach; ISBN 90 5668 059 5	December 1999
61	Mark Voorneveld	Potential Games and Interactive Decisions with Multiple Criteria; ISBN 90 5668 060 9	December 1999
62	Edward Droste	Adaptive Behavior in Economic and Social Environments; ISBN 90 5668 061 7	December 1999
63	Jos Jansen	Essays on Incentives in Regulation and Innovation; ISBN 90 5668 062 5	January 2000
64	Franc J.G.M. Klaassen	Exchange Rates and their Effects on International Trade; ISBN 90 5668 063 3	January 2000
65	Radislav Semenov	Cross-country Differences in Economic Governance: Culture as a major explanatory factor; ISBN 90 5668 065 X	February 2000
66	Alexandre Possajennikov	Learning and Evolution in Games and Oligopoly Models; ISBN 90 5668 064 1	March 2000
67	Eric Gaury	Designing Pull Production Control Systems: Customization and Robustness; ISBN 90 5668 066 8	March 2000
68	Marco Slikker	Decision Making and Cooperation Restrictions; ISBN 90 5668 067 6	April 2000
69	Ruud Brekelmans	Stochastic Models in Risk Theory and Management Accounting; ISBN 90 5668 068 4	June 2000
70	Flip Klijn	A Game Theoretic Approach to Assignment Problems; ISBN 90 5668 069 2	June 2000

No.	Author	Title	Published
71	Maria Montero	Endogenous Coalition Formation and Bargaining; ISBN 90 5668 070 6	June 2000
72	Bas Donkers	Subjective Information in Economic Decision Making; ISBN 90 5668 071 4	June 2000
73	Richard Nahuis	Knowledge and Economic Growth; ISBN 90 5668 072 2	October 2000
74	Kuno Huisman	Technology Investment: A Game Theoretic Real Options Approach; ISBN 90 5668 073 0	October 2000
75	Gijsbert van Lomwel	Essays on Labour Economics; ISBN 90 5668 074 9	November 2000
76	Tina Girndt	Cultural Diversity and Work-Group Performance: Detecting the Rules; ISBN 90 5668 075 7	December 2000
77	Teye A. Marra	The Influence of Proprietary Disclosure Costs on the Decision to Go Public: Theory and Evidence; ISBN 90 5668 076 5	January 2001
78	Xiaodong Gong	Empirical Studies on the Labour Market and on Consumer Demand; ISBN 90 5668 078 1	February 2001
79	Kanat Çamlibel	Complementarity Methods in the Analysis of Piecewise Linear Dynamical Systems; ISBN 90 5668 079 X	May 2001
80	Joost Driessen	Empirical Studies on the Pricing of Bonds and Interest Rate Derivatives; ISBN 90 5668 080 3	June 2001
81	Bram van den Broek	Uncertainty in Differential Games; ISBN 90 5668 081 1	June 2001
82	Ricardo Recht	Organisational Culture and Privatisation: A Case Study of the Argentinean Railway Sector; ISBN 90 5668 082 X	June 2001
83	Willem Verhagen	Inflation Targeting and Interest Rate Policy; ISBN 90 5668 083 8	June 2001
84	Vincent Verouden	Essays in Antitrust Economics; ISBN: 90 5668 086 2	June 2001
85	Rosalia Vazquez Alvarez	A Nonparametric Approach to the Sample Selection Problem in Survey Data; ISBN 90 5668 087 0	June 2001

No.	Author	Title	Published
86	Theo Leers	Public Pensions and Population Ageing: An Economic Analysis of Fertility, Migration, and Social-Security Policy; ISBN 90 5668 084 6	June 2001
87	Bernard Conlon	Consumer Rationality in Choice; ISBN 90 5668 085 4	June 2001
88	Judith Timmer	Cooperative behaviour, uncertainty and operations research; ISBN 90 5668 088 9	September 2001
89	Enrico Diecidue	Nonexpected Utility and Coherence; ISBN 90 5668 089 7	November 2001
90	Alexander Konovalov	Essays in General Equilibrium Theory; ISBN 90 5668 091 9	November 2001
91	Stefan Stremersch	Essays on Marketing Strategy in Technology-Intensive Markets; ISBN 90 5668 090 0	December 2001
92	Rian Drogendijk	Expansion Patterns of Dutch Firms in Central and Eastern Europe: Learning to Internationalize; ISBN 90 5668 092 7	December 2001
93	Han Donker	Takeovers. An Empirical Study of Ownership Structure and the Wealth to Shareholders in Dutch Takeovers; ISBN 90 5668 093 5	December 2001
94	Erica van Herpen	Perceptions and Evaluations of Assortment Variety; ISBN 90 5668 094 3	December 2001
95	Jana Vyrastekova	Cheap Talk and Spiteful Preferences in Ultimatum Games: Experiments and Evolutionary Rationale; ISBN 90 5668 095 1	March 2002
96	Tao Lin	Statistics of Extremes in the Space of Continuous Functions; ISBN 90 5668 096 X	May 2002
97	Sander Onderstal	Papers in Auction Theory; ISBN 90 5668 097 8	May 2002
98	Wolf Wagner	Risk Sharing under Incentive Constraints; ISBN 90 5668 098 6	June 2002
99	Willem-Jan van den Heuvel	Integrating Modern Business Applications with Objectified Legacy Systems; ISBN 90 5668 099 4	June 2002
100	Jacco Wielhouwer	Optimal Tax Depreciation and its Effects on Optimal Firm Investments; ISBN 90 5668 100 1	June 2002
101	Peter Roosenboom	Corporate Governance Mechanisms in IPO Firms; ISBN 90 5668 101 X	September 2002

No.	Author	Title	Published
102	Victoria Shestalova	Essays in Productivity and Efficiency; ISBN 90 5668 103 6	September 2002
103	Paula van Veen-Dirks	Flexibility and Control: An Empirical Study Relating Production Flexibility to the Design of Performance Management Systems; ISBN 90 5668 102 8	September 2002
104	Michal Matějka	Management Accounting in Organizational Design: Three Essays; ISBN 90 5668 105 2	November 2002
105	Maarten Pronk	Market Liquidity around Earnings Announcements; ISBN 90 5668 106 0	November 2002
106	Nancy Kamp-Roelands	Towards a framework for auditing environmental reports; ISBN 90 5668 108 7	November 2002
107	Ana Ferreira	Statistics of extremes: estimation and optimality; ISBN 90 5668 107 9	November 2002
108	Alexei Goriaev	On the behaviour of Mutual Fund Investors and Managers; ISBN 90 5668 110 9	November 2002
109	Paul Ingenbleek	Money for Value Pricing from a Resource-Advantage Perspective; ISBN 90 5668 109 5	December 2002
110	Frank Wijen	Stakeholder influence and organizational learning in environmental management; ISBN 90 5668 111 7	December 2002
111	Michele Belot	Labor market institutions in OECD countries: Origins and consequences; ISBN 90 5668 113 3	January 2003
112	Federica Teppa	Estimating preference parameters in individual financial decision making; ISBN 90 5668 112 5	January 2003
113	Dmitry Danilov	The effects of pretesting in econometrics with applications in finance; ISBN 90 5668 104 4	February 2003

[1] Copies can be ordered from Thela Thesis, Prinseneiland 305, 1013 LP Amsterdam, The Netherlands, phone: + 31 20 6255429; fax: +31 20 6203395; e-mail: office@thelathesis.nl

About the author:
DMITRY DANILOV graduated in Mathematics from St. Petersburg State University, Russia. From 1997 to 2002 he was a graduate student at the Department of Econometrics and CentER at Tilburg University, the Netherlands. His main research interests are pretesting and time-series forecasting, including economic and financial applications.

About the thesis:
The thesis is based on three papers. Each paper investigates different aspects of pretesting in regression analysis. In addition, the introduction contains a review of the literature and a discussion of recent developments. The first paper studies the effects of ignoring the effects of pretesting in estimation. It is shown that not reporting the correct moments leads to a significant distortion of the accuracy of many widely used pretest estimators. The second paper concerns forecasting. It derives the unconditional moments of the one-step-ahead WALS forecast and applies these results to stock market forecasting. The third paper investigates properties of the neutral Laplace WALS estimator in the important case where the variance of the innovations is not known.

ISBN: 90 5668 104 4

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

CentER Dissertation Series